WPKAN-Mixer: Long-Term Time Series Forecasting with Adaptive Function Learning on Wavelet-Decomposed Components

Guoqing Tang[1], Hongwei Zhao[1], Lei Liu[1], Tengyuan Liu[1], Jiahui Huang[1], Bin Li[1,†]
[1]University of Science and Technology of China, Hefei, China
Email:{guoqing_tang,tengyuanliu,jiahuihuang}@mail.ustc.edu.cn
{hwzhao,liulei13,binli}@ustc.edu.cn

***Abstract*—Time series forecasting is widely applied in fields like weather, power, and traffic management, but the complex multi-scale dynamics of real-world sequences make accurate forecasting extremely challenging. Existing methods typically employ effective decomposition techniques for time series forecasting tasks, which simplify complex raw series into more tractable sub-series (e.g., trend, seasonality) for targeted modeling. However, these approaches remain subject to inherent limitations: their modeling units apply a unified representation paradigm to all components, which hinders the adaptive modeling of local nonlinear dynamics across components at multiple scales, thus limiting fitting accuracy. This paper introduces WPKAN-Mixer for long-term time series forecasting, comprising primarily a Wavelet Transform module, a Patching-Embedding module and a dual Kolmogorov-Arnold Networks (KAN) Mixers module. We leverage the time-frequency localization of the wavelet transform module to achieve fine-grained temporal-signal decomposition. Furthermore, we design a dual KAN Mixers module incorporating the spline-based learnable function space to "tailor-make" an optimal functional representation for each frequency component, enabling flexible capture of non-linear dynamics characteristics from both temporal and feature dimensions. Experiments show that WPKAN-Mixer significantly outperforms state-of-the-art models on public benchmarks, achieving an average MSE and MAE reduction of 3-5% on most datasets, while also demonstrating excellent robustness and superior computational efficiency.**

***Index Terms*—Time Series Forecasting, Wavelet Decomposition, Kolmogorov-Arnold Networks, Adaptive Activation Functions, Dual KAN Mixers**

I. INTRODUCTION

Long-Term Time Series Forecasting (LTSF) is a critical task in numerous scientific and engineering domains that aims to learn patterns from complex historical data to predict the distant future, including finance, power management, traffic planning, and weather forecasting [1]–[6]. Its core challenge lies in the fact that real-world time series are often a complex superposition of multiscale dynamics, such as slowly varying annual trends, clear weekly seasonalities, and daily random fluctuations. Therefore, an effective LTSF model must be able to accurately capture and model these coexisting dynamics [1], [7].

Recent advances in LTSF have been largely driven by architectural innovations. On one hand, holistic architectures attempt to capture all dynamics directly with a single powerful model. For instance, Transformer-based models like PatchTST [8]–[10] have improved efficiency in handling long sequences through patching, while the latest State Space Models (SSMs) such as Mamba [11] have emerged as new performance benchmarks with their linear complexity and strong sequence modeling capabilities.

On the other hand, decomposition-based forecasting has emerged as a powerful and parallel paradigm, which simplifies complex raw series into more tractable sub-series, allowing targeted modeling for each component. For example, models like TimeMixer [2] and TSMixer [12] decompose series into trend and seasonal components.

However, a dual limitation persists within these successful decomposition paradigms. At the "Decomposition" stage, mainstream decomposition methods primarily rely on moving averages (for trend-seasonal decomposition) or the Fourier Transform (for periodic decomposition). These tools excel with relatively stationary, periodic data. However, due to their global or fixed nature in the time domain, they struggle to precisely capture and disentangle non-stationary and transient local features, which are often critical for accurate forecasting. At the "Modeling" stage, existing methods often apply a uniform model with a fixed and nonlinear activation function (e.g.an MLP with ReLU or GELU) [2], [12]–[14]—across all time-frequency components. This "one-size-fits-all" approach neglects the fact that smooth low-frequency trends and highly variable high-frequency noises exhibit fundamentally different dynamics, which requires locally modulating temporal and feature characteristics, leading to limited expressiveness.

To address the aforementioned dual limitations, this paper proposes WPKAN-Mixer, a novel long-term time series forecasting architecture that integrates fine-grained wavelet-domain decomposition with learnable function spaces to achieve accurate prediction of complex non-stationary sequences. Specifically, we employ a Wavelet Transform [13], [15] module to leverage its superior time-frequency localization, enabling precise multiscale decomposition that captures non-stationary and transient features. Building on this, since recent studies have shown that Kolmogorov-Arnold Networks (KAN) [16]–[18] exhibit greater potential than MLPs in efficiency, expressivity and interpretability [19]–[21], we introduce a dual KAN Mixers module, which equips each

†: Corresponding author.

frequency component with a spline-based learnable function space for adaptive modeling. This design allows the model to flexibly capture nonlinear dynamics across different frequency components while separately modulating temporal evolution and inter-feature interactions. By combining wavelet-based disentanglement with component-wise adaptive modeling, WPKAN-Mixer can "tailor-make" optimal functional representations for each sub-series, effectively modeling both smooth low-frequency trends and highly variable high-frequency fluctuations, thereby providing a principled framework for precise and expressive long-term forecasting of complex time series.

The main contributions of this paper are summarized as follows:

- We design WPKAN-Mixer, a novel architecture that integrates wavelet decomposition with patching-embedding and dual KAN mixers to precisely separate multiscale non-stationary dynamics and adaptively capture nonlinear interactions in both temporal and feature dimensions.
- We further assign spline-based learnable functions to each decomposed frequency component, enabling component-specific nonlinear mappings for fine-grained modeling of both low- and high-frequency dynamics.
- Experiments demonstrate that WPKAN-Mixer significantly outperforms existing SOTA models, achieving an average MSE and MAE improvement of 3-5% in long-term forecasting tasks, validating our paradigm's effectiveness, while maintaining excellent robustness in terms of computational efficiency and hyperparameter sensitivity.

For reproducibility, the source code is available at:https://anonymous.4open.science/r/WPKAN-Mixer-6F04.

II. RELATED WORK

The recent literature on Long-Term Time Series Forecasting is dominated by two main paradigms: holistic models that perform end-to-end modeling on the raw sequence, and decomposition-based models that follow a "divide and conquer" strategy. In parallel, innovations in fundamental neural network building blocks, like Kolmogorov-Arnold Networks (KANs), offer new perspectives for LTSF.

A. Holistic Models for Time Series Forecasting

Holistic models aim to learn complex temporal patterns directly from the raw, mixed-dynamics sequence using a single, powerful architecture. Transformers [9], [10], [22]–[24] were a dominant force, but were limited by their quadratic complexity. Subsequent works like PatchTST [8] achieved linear complexity via patching. More recently, State Space Models (SSMs) [11], particularly Mamba, have set new performance benchmarks with their linear-time efficiency. While powerful, these models require a strong implicit capacity to disentangle dynamics.

B. Decomposition-based Time Series Forecasting

In contrast to holistic models, decomposition-based models simplify the problem by explicitly disentangling the series into more tractable components. Mainstream methods include trend-seasonal decomposition via moving averages (e.g., DLinear [7], TimeMixer [2], [12]–[14]) and frequency-domain decomposition via Fourier Transform. However, these tools have limited ability to capture non-stationary and transient local features. The Wavelet Transform [13], [15] offers a superior solution due to its excellent time-frequency localization, with its potential validated in works like WPMixer.

C. Kolmogorov-Arnold Networks

Proposed as a fundamental alternative to Multi-Layer Perceptrons (MLPs), the recently introduced Kolmogorov-Arnold Networks [16]–[18] are inspired by the Kolmogorov-Arnold representation theorem and have initiated a paradigm shift in the basic construction of neural networks. Unlike MLPs, which concentrate their learning capacity on the linear weights connecting nodes, KANs place learnable activation functions on the edges of the computational graph. These activation functions are parameterized by B-splines and can adaptively learn their functional shape based on the data.

This core design endows KANs with significant advantages over traditional MLPs in terms of expressivity, parameter efficiency, and interpretability. The advent of KANs has spurred rapid exploration of their application in various domains, including signal processing and time series, with initial works like Wav-KAN [17] and TKAN [21] already emerging. Our work distinguishes itself from these other explorations by being the first to design and employ Dual KAN Mixers for token-mixing and channel-mixing within a wavelet decomposition framework, enabling the flexible capture of nonlinear dynamics characteristics from both temporal and feature dimensions.

III. METHODOLOGY

WPKAN-Mixer employs a "Decomposition, Forecasting and Reconstruction" paradigm, driven by our core motivation to match adaptive functions with multiscale time series dynamics, which is specifically demonstrated in Figure 1. To mitigate distribution shifts common in time series, the model is framed by Reversible Instance Normalization (RevIN) [25] to improve generalization. The three core stages of our approach are as follows:

A. Multi-level Wavelet Decomposition

To achieve an effective disentanglement of mixed dynamics, we employ an m-level Discrete Wavelet Transform (DWT) [15], [26]. DWT is a powerful signal processing tool that iteratively decomposes a time series into sub-series of different frequencies. Given a RevIN-normalized input sequence $X_L \in \mathbb{R}^{L \times C}$, where L is the look-back window length and C is the number of variates (channels), the DWT decomposes it using a specific wavelet basis (e.g., Daubechies, Coiflets, denoted as ψ).

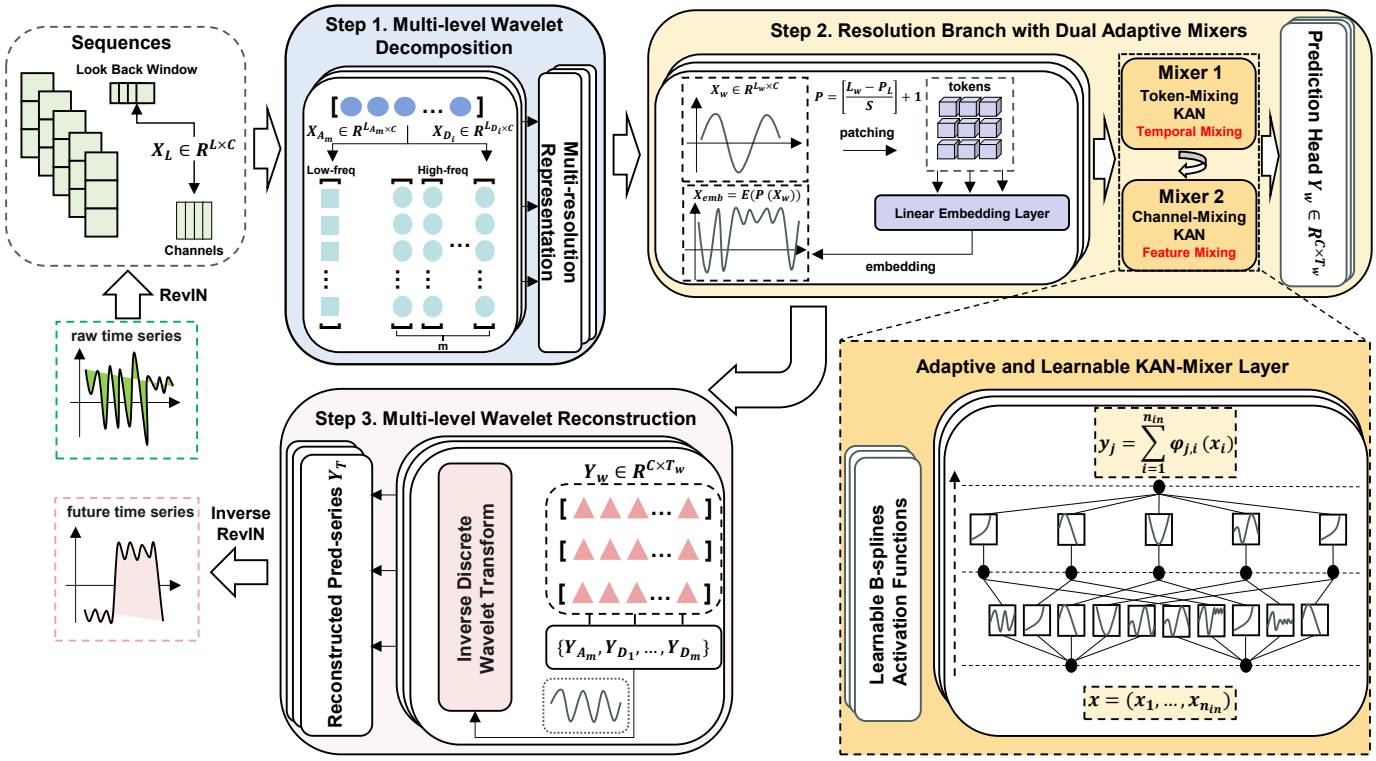

Fig. 1. Overall Diagram of the WPKAN-Mixer. The model first decomposes the input time series into multiple frequency components using a Multi-level Discrete Wavelet Transform (DWT). The core innovation is **fusing each wavelet-decomposed component with function-adaptive matching for the first time**, independently processed by **a KAN-based resolution branch, where the KAN-Mixer adaptively learns an optimal functional form for it**. This is achieved by using **two sequential KAN-Mixer layers for token-mixing and channel-mixing**. Finally, all predicted components are reconstructed into the final output via an Inverse Wavelet Transform (IDWT).

$$[X_{Am}, X_{Dm}, \ldots, X_{D1}] = \text{DWT}(X_L, \psi, m) \tag{1}$$

After decomposition, we obtain one approximation coefficient series $X_{Am} \in \mathbb{R}^{L_{Am} \times C}$ and m detail coefficient series $X_{Di} \in \mathbb{R}^{L_{Di} \times C}$. The approximation coefficients X_{Am} represent the primary low-frequency information of the series, i.e., the long-term trend, while the detail coefficients X_{Di} capture the high-frequency information at the i-th scale, such as short-term fluctuations and noise. These $m+1$ coefficient series together form a multi-resolution representation of the original series, laying the foundation for the subsequent targeted modeling.

B. Resolution Branch with Dual Adaptive Mixers

This is the technical vehicle for realizing our core motivation. Each of the $m+1$ decomposed series is fed into an independent, identically structured resolution branch. By isolating modeling across frequency scales, the architecture prevents interference between low- and high-frequency components, enabling the model to specialize and capture scale-specific patterns more effectively [16], [19].

1. Patching and Embedding: Within each branch, the input coefficient series $X_w \in \mathbb{R}^{L_w \times C}$ is first "patched" [8], [27]. We segment the sequence of length L_w into P overlapping patches of length P_L with a stride of S. The number of patches is $P = \lfloor (L_w - P_L)/S \rfloor + 1$. This process encapsulates local temporal information into discrete "tokens". Subsequently, a shared linear embedding layer E projects each patch into a d_{model}-dimensional feature space:

$$X_{emb} = E(\text{Patching}(X_w)) \in \mathbb{R}^{C \times P \times d_{\text{model}}} \tag{2}$$

2. KAN-Mixer Block: This module is the core of WPKAN-Mixer, implementing information interaction across the temporal and feature dimensions of the patched representation via two KAN layers [16], [17], [19]. The fundamental operational unit of a KAN layer is a learnable activation function. For an input vector $\mathbf{x} = (x_1, \ldots, x_{n_{in}})$, the output y_j of a KAN layer is defined as:

$$y_j = \sum_{i=1}^{n_{in}} \phi_{j,i}(x_i), \quad j = 1, \ldots, n_{out} \tag{3}$$

where $\phi_{j,i}$ is a learnable univariate function parameterized by B-splines. This contrasts with the fixed activation(linear-transform) pattern in MLPs, endowing KAN with superior flexibility in function approximation.

(1) Token-Mixing KAN: To capture dependencies among different temporal patches, we first mix along the patch dimension. The input X_{emb} is passed through a LayerNorm and a permutation operator τ before being fed into the token-

mixing KAN layer. We use a residual connection to stabilize training:

$$\hat{X}_{emb} = \text{LayerNorm}(X_{emb}) \tag{4}$$

$$X_{tok} = X_{emb} + \mathcal{T}^{-1}(\text{KAN}_{\text{token}}(\mathcal{T}(\hat{X}_{emb}))) \tag{5}$$

The learnable spline functions ψ within KAN_{token} adaptively learn the optimal temporal mixing patterns from the data.

(2) Channel-Mixing KAN: Next, we mix along the feature dimension to learn interactions among different features within each time patch.

$$\hat{X}_{tok} = \text{LayerNorm}(X_{tok}) \tag{6}$$

$$X_{out} = X_{tok} + \text{KAN}_{\text{channel}}(\hat{X}_{tok}) \tag{7}$$

The spline functions in $KAN_{channel}$ adaptively learn the nonlinear relationships between features.

3. Prediction Head: After processing through several KAN-Mixer blocks, the output tensor X_{out} is first flattened and then projected by a linear layer to the required prediction length T_w for that branch, obtaining the predicted coefficient series $Y_w \in \mathbb{R}^{C \times T_w}$.

C. Multi-level Wavelet Reconstruction

Finally, all predicted coefficient series from the resolution branches, $\{Y_{Am}, Y_{D1}, \ldots, Y_{Dm}\}$, are precisely reconstructed into the final prediction result via the Inverse Discrete Wavelet Transform (IDWT) [15], [28], yielding the intermediate prediction $\hat{Y}_T$:

$$\hat{Y}_T = \text{IDWT}([Y_{Am}, Y_{Dm}, \ldots, Y_{D1}]) \in \mathbb{R}^{T \times C} \tag{8}$$

where $\hat{Y}_T$ is the reconstructed series in the normalized space. The final output Y_T is obtained by applying the reverse RevIN transformation to $\hat{Y}_T$ to restore its original numerical scale.

IV. EXPERIMENTS

A. Experimental Setup

Datasets: We evaluated our model on 7 widely used public datasets, including ETTh1, ETTh2, ETTm1, ETTm2, weather, electricity and traffic [29], [30] all specified in Table I.

TABLE I
SPECIFICATIONS OF THE DATASETS.

Dataset	Variates	Timesteps	Frequency
ETTh1, ETTh2	7	17,420	1 hour
ETTm1, ETTm2	7	69,680	15 min
Weather	21	52,696	10 min
Electricity	321	26,304	1 hour
Traffic	862	17,544	1 hour

Baselines: To comprehensively evaluate WPKAN-Mixer, we compare it against a diverse set of state-of-the-art baselines covering the main technical paradigms. These include MLP-based models like TimeMixer [2], WPMixer [13] and CNN-based model TimesNet [31]; efficient Transformer-based models PatchTST [8] and the cross-variate focused Crossformer [23]; as well as architectures leveraging convolutional and frequency-domain techniques, MICN [32] and FiLM [33], respectively.

TABLE II
UNIFIED HYPERPARAMETER ON EACH DATASET. BS AND GS REPRESENT BATCH SIZE AND GRADIENT ACCUMULATION STEPS RESPECTIVELY.

Dataset	Look back window	BS / GS	Epoch	Patience
ETT	96	64/2	10	5
Weather	96	64/2	20	5
Electricity	96	8/4	20	10
Traffic	96	4/2	20	10

Implementation Details: For a rigorous and fair comparison, we follow the standard protocol of baseline experiments [1]. Each experiment was performed with a consistent batch size, epoch, and look back window(Table II). We used Mean Squared Error (MSE) and Mean Absolute Error (MAE) for the evaluation. All experiments were carried out on NVIDIA GeForce RTX 4090 GPUs.

For the key hyperparameters of our model, we conducted a comprehensive hyperparameter search for each dataset using the Optuna framework to determine the optimal configuration. Table IV details the search space used for the 200 trials conducted on the ETTh1 dataset. For the KAN layers, we adopt a consistent grid configuration and spline degree across all datasets to balance computational efficiency with non-linear approximation capacity. The final hyperparameter combination used in our main experiments represents the best-performing configuration on the validation set.

All reported MSE and MAE results are averaged over 3 independent runs with different random seeds to minimize the impact of stochastic variance.

B. Main Results

As shown in Table III, WPKAN-Mixer demonstrates comprehensive state-of-the-art performance. Based on the aggregated 1st Count metrics, WPKAN-Mixer achieves the top result in a total of 53 settings (24 for MSE, 29 for MAE), significantly ahead of the second-best model WPMixer and far surpassing all other baselines.

This superiority is most evident when compared to its direct predecessor, WPMixer. On the four ETT datasets, which are known for their complex non-stationary dynamics, WPKAN-Mixer consistently outperforms its MLP-based counterpart. For example, on the ETTh1 dataset, our model achieves an average MSE reduction of 3.3%. The performance gain is particularly pronounced at longer horizons, such as the 4.9% MSE reduction on ETTh1 for a 720-step forecast, demonstrating the

TABLE III

MULTIVARIATE LONG-TERM FORECASTING RESULTS. **SMALLER MSE AND MAE INDICATE HIGHER PREDICTION ACCURACY. BOLD TEXT** REPRESENTS THE OPTIMAL, UNDERLINED TEXT REPRESENTS THE SUBOPTIMAL. TO ENSURE A FAIR COMPARISON, ALL BASELINE MODELS WERE IMPLEMENTED USING THEIR OFFICIAL OPEN-SOURCE CODE WITH RECOMMENDED HYPERPARAMETERS.

Dataset	Horizon	WPKAN-Mixer (Ours)		WPMixer (2025)		TimeMixer (2024)		TimesNet (2023)		PatchTST (2023)		MICN (2023)		Crossformer (2023)		FiLM (2022a)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.366**	**0.387**	0.372	0.393	0.398	0.403	0.434	0.444	0.379	0.399	0.438	0.444	0.406	0.437	0.522	0.493
	192	**0.419**	**0.416**	0.427	0.417	0.440	0.430	0.488	0.474	0.426	0.432	0.442	0.454	0.458	0.457	0.572	0.521
	336	**0.445**	**0.431**	0.465	0.434	0.501	0.460	0.523	0.487	0.470	0.458	0.597	0.552	0.614	0.577	0.627	0.549
	720	**0.429**	**0.440**	0.451	0.447	0.497	0.480	0.528	0.503	0.511	0.500	0.839	0.684	0.762	0.648	0.704	0.607
	Avg	**0.415**	**0.418**	0.429	0.423	0.459	0.443	0.493	0.477	0.447	0.447	0.579	0.534	0.560	0.530	0.606	0.543
ETTh2	96	**0.279**	**0.332**	0.285	0.337	0.299	0.345	0.339	0.374	0.302	0.364	0.389	0.426	0.694	0.576	0.351	0.386
	192	**0.353**	**0.381**	0.366	0.387	0.385	0.407	0.397	0.410	0.376	0.414	0.521	0.496	1.088	0.724	0.428	0.430
	336	**0.375**	0.408	0.379	**0.407**	0.430	0.439	0.468	0.463	0.424	0.445	0.582	0.534	1.247	0.805	0.450	0.456
	720	**0.403**	**0.428**	0.419	0.436	0.444	0.453	0.496	0.485	0.473	0.484	0.806	0.645	1.489	0.885	0.458	0.465
	Avg	**0.352**	**0.387**	0.362	0.392	0.390	0.411	0.425	0.433	0.394	0.427	0.575	0.525	1.130	0.748	0.422	0.434
ETTm1	96	**0.309**	**0.346**	0.317	0.351	0.323	0.360	0.338	0.378	0.327	0.366	0.319	0.372	0.360	0.407	0.408	0.418
	192	**0.354**	**0.373**	0.363	0.377	0.375	0.390	0.382	0.399	0.368	0.390	0.363	0.398	0.382	0.417	0.449	0.438
	336	**0.380**	**0.394**	0.388	0.397	0.394	0.407	0.424	0.428	0.399	0.409	0.385	0.414	0.625	0.586	0.481	0.458
	720	**0.441**	**0.432**	0.449	0.433	0.457	0.441	0.478	0.456	0.458	0.445	0.516	0.504	0.736	0.653	0.546	0.493
	Avg	**0.371**	**0.386**	0.379	0.390	0.387	0.400	0.406	0.415	0.388	0.403	0.396	0.422	0.526	0.516	0.471	0.452
ETTm2	96	**0.170**	**0.252**	0.171	0.253	0.175	0.256	0.185	0.265	0.184	0.268	0.189	0.286	0.274	0.362	0.194	0.275
	192	**0.235**	0.296	**0.235**	**0.295**	0.238	0.300	0.252	0.309	0.247	0.306	0.285	0.358	0.621	0.597	0.258	0.314
	336	**0.292**	**0.334**	0.293	**0.334**	0.297	0.337	0.317	0.347	0.310	0.348	0.396	0.431	1.407	0.840	0.318	0.351
	720	**0.387**	0.391	0.387	**0.390**	0.395	0.395	0.416	0.405	0.421	0.415	0.544	0.514	3.114	1.240	0.417	0.406
	Avg	**0.271**	**0.318**	0.272	**0.318**	0.276	0.322	0.293	0.332	0.291	0.334	0.354	0.397	1.354	0.760	0.297	0.337
Electricity	96	**0.147**	**0.240**	0.150	0.241	0.156	0.247	0.168	0.272	0.190	0.296	0.165	0.276	0.153	0.256	0.407	0.459
	192	0.164	0.254	**0.163**	**0.253**	0.170	0.261	0.186	0.288	0.199	0.304	0.176	0.288	0.171	0.270	0.383	0.440
	336	**0.177**	**0.268**	0.179	0.270	0.188	0.278	0.203	0.305	0.217	0.320	0.186	0.298	0.203	0.304	0.417	0.462
	720	0.223	**0.302**	**0.220**	0.305	0.228	0.313	0.223	0.320	0.258	0.352	0.208	0.316	0.252	0.340	0.475	0.509
	Avg	**0.178**	**0.266**	**0.178**	0.267	0.186	0.275	0.195	0.296	0.216	0.318	0.184	0.295	0.195	0.293	0.421	0.468
Weather	96	**0.160**	**0.204**	0.163	0.206	0.162	0.209	0.175	0.226	0.174	0.217	0.192	0.250	0.174	0.239	0.202	0.242
	192	0.212	0.249	0.209	**0.247**	**0.207**	0.251	0.235	0.274	0.223	0.257	0.240	0.301	0.232	0.301	0.245	0.276
	336	0.266	0.290	**0.263**	**0.288**	**0.263**	0.291	0.283	0.306	0.280	0.298	0.282	0.331	0.277	0.340	0.294	0.310
	720	0.344	0.339	**0.340**	**0.338**	0.345	0.344	0.360	0.355	0.356	0.349	0.350	0.387	0.372	0.412	0.365	0.354
	Avg	0.245	**0.270**	**0.244**	**0.270**	**0.244**	0.274	0.263	0.290	0.258	0.280	0.266	0.317	0.264	0.323	0.277	0.296
Traffic	96	0.481	**0.281**	0.466	0.286	0.469	0.294	0.604	0.320	**0.459**	0.298	0.522	0.314	0.644	0.429	0.647	0.384
	192	0.483	**0.288**	0.475	0.290	0.513	0.302	0.625	0.325	**0.469**	0.301	0.536	0.317	0.665	0.431	0.600	0.362
	336	0.512	**0.285**	0.488	**0.297**	0.514	0.309	0.634	0.335	**0.483**	0.307	0.550	0.322	0.674	0.420	0.611	0.368
	720	0.536	**0.310**	0.519	0.313	0.550	0.322	0.664	0.347	**0.517**	0.326	0.576	0.333	0.683	0.424	0.695	0.427
	Avg	0.503	**0.291**	0.487	0.297	0.512	0.307	0.632	0.332	**0.482**	0.308	0.546	0.322	0.667	0.426	0.638	0.385
1st Count:		24	29	7	11	3	0	0	0	5	0	0	0	0	0	0	0

advantage of our adaptive modeling in capturing long-term dependencies.

Furthermore, WPKAN-Mixer establishes a new SOTA by outperforming other leading paradigms. For example, compared to the strong Transformer-based baseline PatchTST, our model reduces the average MAE on the Weather dataset by 3.6%. While WPMixer and PatchTST show competitive performance on the Traffic dataset, which may be attributed to its specific signal characteristics, WPKAN-Mixer's overwhelming overall performance across the broader and more diverse set of benchmarks strongly validates our core motivation: combining the fine-grained disentanglement of wavelets with the adaptive function learning of KAN is the key to achieving more accurate and robust long-term time series forecasting.

C. Efficiency Analysis

To comprehensively evaluate the trade-offs between performance, computational overhead, and memory footprint of WPKAN-Mixer, we provide a visual comparison against several key baselines in Figure 2. For a fair comparison, all models were evaluated under a unified lightweight configuration (hidden dimension to $d_{\text{model}} = 16$, batch size to 64, look back window to $L = 96$).

The analysis of Figure 2 leads to a clear conclusion: WPKAN-Mixer consistently occupies the bottom-left quadrant on both datasets, indicating an exceptional balance across three critical dimensions. It achieves the lowest MSE (0.377 on ETTh1, 0.165 on Weather), demonstrating the highest accuracy. Crucially, this is achieved with high efficiency; its training speed is on par with lightweight models like TimeMixer and FiLM, while its minimal memory footprint (approx. 1GB) is substantially smaller than memory-intensive models like PatchTST and TimesNet. In summary, WPKAN-Mixer demonstrates a Pareto-optimal balance, delivering state-of-the-art accuracy with a highly competitive computational budget and validating its value as an efficient, powerful, and practical LTSF solution.

TABLE IV
HYPERPARAMETER SEARCH SPACE FOR WPKAN-MIXER ON ETTH1 DATASET.

Hyperparameter	**Search Space / Value**
Learning Rate (η)	Log-uniform [1e-5, 1e-2]
Hidden Dimension (d_{model})	{128,256}
KAN Expansion Factor (f_t, f_d)	{3, 5, 7, 8}
Patch Length / Stride	{12, 16} / {4, 8}
Dropout Rate	{0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5}
Embedding Dropout	{0.0, 0.05, 0.1, 0.2, 0.4}
Weight Decay	{0.0}
LR Scheduler ('lradj')	{type1, type2, type3}
Wavelet Basis (ψ)	{db2, db3, db5, sym2-5, coif1, coif4-5, bior3.1}
Decomposition Level (m)	{1, 2, 3}

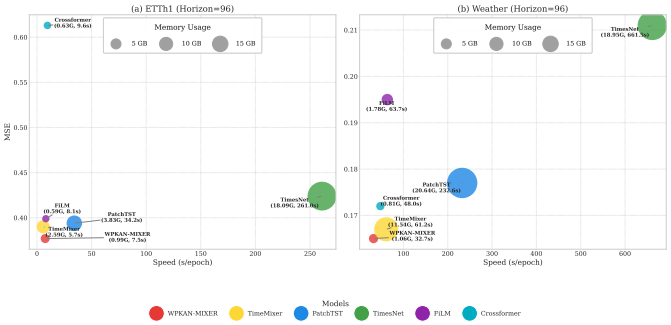

Fig. 2. Model Performance Comparison. Subfigures (a) and (b) illustrate the performance of different models on ETTh1 and Weather datasets, respectively. Larger spheres indicate greater memory usage.

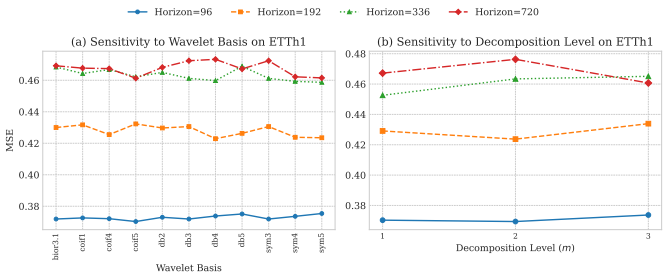

Fig. 3. Sensitivity Analysis on ETTh1.The figure illustrates the model's MSE performance on the ETTh1 dataset across different horizons as the wavelet basis (a) and decomposition level (b) are varied.

D. Sensitivity Analysis

To investigate the model's robustness to key hyperparameters, we conducted a sensitivity analysis on the ETTh1 dataset by varying the wavelet basis and decomposition level, with results visualized in Figure 3. We observe high stability in performance across all 11 different wavelet bases tested, as the MSE values fluctuate within a very narrow range for all horizons (Fig. 3(a)). For the decomposition level (Fig. 3(b)), the model also exhibits stable performance, with a level of m=2 or m=3 generally offering a slight improvement. In summary, this analysis strongly validates the robustness of WPKAN-Mixer: while marginal gains can be obtained through hyperparameter optimization, its core performance is not critically dependent on these specific design choices.

E. Ablation Study

To thoroughly dissect the contributions of each component within WPKAN-Mixer, we designed a series of fine-grained ablation studies, with results presented in Table V. We quantify their impact on model performance by removing key modules or altering their core implementation.

The results in Table V show that our full model performs best. Removing any core module leads to significant performance degradation, proving our overall architectural design is sound. Most importantly, when switching the implementation from KAN to MLP or removing wavelet decomposition, there is a consistent and clear drop in performance. This irrefutably validates our core motivation that combining well-disentangled wavelet transform with adaptive function learning of KAN is the key factor in achieving the performance breakthrough.

TABLE V
ABLATION OF EACH MODULE IN WPKAN-MIXER ON ETTH1, ETTH2, ETTM1 AND ETTM2 WITH MSE. D, T_x, C_x, K_a REPRESENT THE WAVELET DECOMPOSITION MODULE, TOKEN MIXING LAYER, CHANNEL MIXING LAYER, KAN-LEARNABLE ACTIVATION.

Case	Modules				MSE			
	D	T_x	C_x	K_a	ETTh1	ETTh2	ETTm1	ETTm2
I	✓	✓	✓	✓	**0.415**	**0.352**	**0.371**	**0.271**
II	✗	✓	✓	✓	0.440	0.365	0.381	0.277
III	✓	✗	✓	✓	0.431	0.362	0.380	0.283
IV	✓	✓	✗	✓	0.439	0.364	0.388	0.277
V	✓	✓	✓	✗	0.429	0.362	0.379	0.275

V. CONCLUSION

In this paper, we propose WPKAN-Mixer, a novel wavelet-KAN architecture that overcomes the limitations of coarse

decomposition and fixed nonlinear modeling in existing forecasting methods. The model leverages the Wavelet Transform module for fine-grained time-frequency disentanglement of non-stationary components and employs dual KAN mixers with spline-based learnable activations to adaptively model non-linear dynamics characteristics across both temporal and feature dimensions. Experiments on benchmark datasets demonstrate that WPKAN-Mixer consistently outperforms state-of-the-art methods, achieving 3–5% improvement in MSE and MAE while maintaining a highly competitive computational budget and demonstrating strong robustness to hyperparameter choices. Our work validates that the synergy of precise multiscale decomposition and adaptive function learning is a more principled path toward modeling complex dynamics in long-term time series forecasting.

ACKNOWLEDGMENT

This research is supported by the Key Science & Technology Project of Anhui Province (No. 202523F12050009), the Fundamental Research Funds for the Central Universities (Grant No.WK2080000206), and the Opening Project of the State Key Laboratory of General Artificial Intelligence (Project No.SKLAGI2025OP06).

REFERENCES

[1] Y. Wang, H. Wu, J. Dong, Y. Liu, M. Long, and J. Wang, "Deep time series models: A comprehensive survey and benchmark," *unpublished*, 2024. arXiv preprint arXiv:2407.13278.

[2] S. Wang, H. Wu, X. Shi, T. Hu, H. Luo, L. Ma, J. Y. Zhang, and J. ZHOU, "Timemixer: Decomposable multiscale mixing for time series forecasting," in *International Conference on Learning Representations (ICLR)*, 2024.

[3] D. S, P. S. Reddy, M. Vyshnevy, K. Tejasri, B. Akshaya, and P. Nikhil, "Prediction and analysis of stock markets using arima models," in *2024 3rd International Conference on Computational Modelling, Simulation and Optimization (ICCMSO)*, pp. 268–271, 2024.

[4] Y. Kumar Jain, C. Shekhar Upadhyay, C. Mahadeva Murthy, M. T. Bhagawati, K. P. Veena, and A. Mandal Sarkar, "Advanced machine learning techniques for forecasting financial market trends and volatility," in *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, pp. 1–6, 2024.

[5] Z. Wang and Y. Yang, "Tf-dpnet: A time-frequency with dual perspectives network for multivariate time series prediction," in *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2025.

[6] D. Zhang, Z. Zhang, and Y. Wang, "Multivariate long sequence time series forecasting based on robust spatiotemporal attention," in *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024.

[7] A. Zeng, M.-H. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?," in *AAAI Conference on Artificial Intelligence*, 2022.

[8] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations*, 2023.

[9] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, *et al.*, "Transformers in time series: A survey," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 6778–6786, 2023.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *unpublished*, 2024. arXiv preprint arXiv:2312.00752.

[12] S.-A. Chen, C.-L. Li, S. O. Arik, N. C. Yoder, and T. Pfister, "TSMixer: An all-MLP architecture for time series forecast-ing," *Transactions on Machine Learning Research*, 2023.

[13] M. M. N. Murad, M. Aktukmak, and Y. Yilmaz, "WPMixer: Efficient multi-resolution mixing for long-term time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 19581–19588, Apr. 2025.

[14] A. Das, W. Kong, A. Leach, S. K. Mathur, R. Sen, and R. Yu, "Long-term forecasting with tiDE: Time-series dense encoder," *Transactions on Machine Learning Research*, 2023.

[15] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks," *ArXiv*, vol. abs/1805.08620, 2018.

[16] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljacic, *et al.*, "KAN: Kolmogorov–Arnold networks," in *International Conference on Learning Representations (ICLR)*, vol. 2025, pp. 70367–70413, 2025.

[17] Z. Bozorgasl and H. Chen, "Wav-kan: Wavelet kolmogorov-arnold networks," *arXiv preprint arXiv:2405.12832*, 2024.

[18] J. Xu, Z. Chen, J. Li, S. Yang, W. Wang, X. Hu, and E. Ngai, "Enhancing graph collaborative filtering with fourierkan feature transformation," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5376–5380, 2025.

[19] Y.-C. Hong, B. Xiao, and Y. Chen, "TSKANMixer: Kolmogorov-Arnold networks with MLP-Mixer model for time series forecasting," *unpublished*, 2025. arXiv preprint arXiv:2502.18410.

[20] A. D. Bodner, A. S. Tepsich, J. N. Spolski, and S. Pourteau, "Convolutional Kolmogorov-Arnold networks," *unpublished*, 2025. arXiv preprint arXiv:2406.13155.

[21] R. Genet and H. Inzirillo, "Tkan: Temporal kolmogorov-arnold networks," *arXiv preprint arXiv:2405.07344*, 2024.

[22] K. Yang, X. Wang, Z. Wang, C. Chi, C. Deng, and J. Feng, "Stnet: Spatial-temporal transformers are effective for multivariate time series forecasting," in *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2025.

[23] 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 *International Conference on Learning Representations, ICLR*, 2022.

[24] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," 2022.

[25] T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo, "Reversible instance normalization for accurate time-series forecasting against distribution shift," in *International Conference on Learning Representations*, 2022.

[26] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[27] P. Tang and W. Zhang, "Unlocking the power of patch: Patch-based MLP for long-term time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 12640–12648, Apr. 2025.

[28] P. Liu, H. Zhang, W. Lian, and W. Zuo, "Multi-level wavelet convolutional neural networks," *IEEE Access*, vol. 7, pp. 74973–74985, 2019.

[29] M. Štěpnička and M. Burda, "On the results and observations of the time series forecasting competition cif 2016," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, 2017.

[30] S. Ben Taieb, G. Bontempi, A. F. Atiya, and A. Sorjamaa, "A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7067–7083, 2012.

[31] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations*, 2023.

[32] 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 *The Eleventh International Conference on Learning Representations*, 2023.

[33] 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 *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 12677–12690, Curran Associates, Inc., 2022.