MSTN: A Lightweight and Fast Model for General Time-Series Analysis

Anonymous authors
Paper under double-blind review

Abstract

Real-world time series often exhibit strong non-stationarity, complex nonlinear dynamics, and behavior expressed across multiple temporal scales, from rapid local fluctuations to slow-evolving long-range trends. However, many contemporary architectures impose rigid, fixed-scale structural priors—such as patch-based tokenization, predefined receptive fields, or frozen backbone encoders—which can over-regularize temporal dynamics and limit adaptability to abrupt high-magnitude events. To handle this, we introduce the *Multi-scale Temporal Network* (MSTN), a hybrid neural architecture grounded in an *Early Temporal Aggregation* principle. MSTN integrates three complementary components: (i) a multi-scale convolutional encoder that captures fine-grained local structure; (ii) a sequence modeling module that learns long-range dependencies through either recurrent or attention-based mechanisms; and (iii) a self-gated fusion stage incorporating squeeze–excitation and a single dense layer to dynamically reweight and fuse multi-scale representations. This design enables MSTN to flexibly model temporal patterns spanning milliseconds to extended horizons, while avoiding the computational burden typically associated with long-context models. Across extensive benchmarks covering imputation, long term forecasting, short term forecasting, classification, and cross-dataset generalization, MSTN achieves state-of-the-art performance, establishing new best results on **33 of 40** datasets, while remaining lightweight ($\sim$278,520 params for MSTN-BiLSTM and $\sim$950,776 $\approx$ 1M for MSTN-Transformer) and suitable for low-latency inference **(<1 sec, often in milliseconds)**, resource-constrained deployment.

1 Introduction

Multivariate time series (MTS) analysis underpins a broad spectrum of societally and industrially critical applications, including healthcare monitoring (Morid et al., 2023), intelligent transportation systems (Kadiyala & Kumar, 2014), climate and environmental modeling (Gruca et al., 2022), energy systems (Kardakos et al., 2013), industrial prognostics, and behavioural analytics. Across these domains, learning from temporal data is central to tasks such as long-term forecasting, missing-value imputation, and trajectory or activity classification (Wu et al., 2022; Zhao et al., 2024). Despite decades of study, robust temporal modeling remains fundamentally challenging. Unlike language or vision, individual time points in a time series carry limited semantic meaning; useful information emerges only through temporal variation patterns such as continuity, periodicity, and long-term trends (Wu et al., 2023; Lim & Zohren, 2021). This intrinsic dependency structure renders time series modeling both information-rich and computationally demanding.

Deep learning (DL) has substantially advanced time series analysis, not only improving predictive accuracy but also enabling transferable representations for downstream tasks (Nie et al., 2023). However, recent evidence that simple linear models can rival or even surpass sophisticated Transformer-based architectures on standard benchmarks (Zeng et al., 2022) has exposed structural limitations in contemporary designs. Classical neural architectures—including convolutional networks (Franceschi et al., 2019; Bai et al., 2018) and recurrent models with gating mechanisms (Hochreiter & Schmidhuber, 1997; Lai et al., 2018)—are constrained by vanishing gradients, limited parallelism, or restricted receptive fields. Although temporal convolutional networks improve efficiency (He & Zhao, 2019), they rely on fixed receptive scales. Transformer-based models promise global dependency modeling (Vaswani et al., 2023), yet their quadratic complexity in

sequence length has driven a proliferation of sparsification and approximation strategies (Zhou et al., 2021; 2022). Many such methods continue to treat time steps as independent tokens, limiting their ability to encode local semantic structure and multi-scale temporal interactions.

These design pressures have led to the emergence of three dominant modeling paradigms, formalized as Channel Strategies (Qiu et al., 2025): Channel Independence (CI), Channel Dependence (CD), and Channel Partiality (CP). CI-based approaches prioritize computational efficiency by processing each variable independently, but sacrifice inter-variable interactions. CD-based models explicitly capture cross-channel dependencies at the cost of poor scalability. CP-based methods seek a compromise by projecting channels into lower-dimensional latent spaces, improving efficiency while retaining partial correlations. Despite their differences, all three paradigms rely on rigid structural assumptions—fixed patch sizes, predefined resolutions, uniform-scale mixing MLPs, static backbones, or explicit periodic decompositions—which limit adaptability to non-stationary, irregular, and long-sequence dynamics. As a result, even state-of-the-art (SOTA) systems such as PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023), SOFTS (Han et al., 2024a), MCformer (Han et al., 2024b), TSMixer (Chen et al., 2023), and recent LLM-inspired approaches (Chang et al., 2025; Jin et al., 2024) struggle to maintain temporal fidelity under complex, real-world conditions.

A key limitation shared by these approaches is the implicit assumption of a single dominant temporal scale. This limitation is architecturally evident in designs such as TSMixer, where the time-mixing MLPs are shared across all time steps, resulting in a fixed-resolution, uniform transformation across the temporal dimension. Real-world temporal processes, however, are inherently multi-scale: micro-level fluctuations, sudden spikes, intermediate dynamics, and long-term trends often coexist and interact. Capturing such a hierarchical structure is essential in safety-critical and high-variability domains, including physiological monitoring, risky driving behavior analysis, and mechanical prognostics. Consequently, fixed-scale or uniform-scale architectures are ill-suited to these settings. Existing models handle a single task or focus on a single type of temporal variations- short or long, rarely balancing both effectively. There is a need for architectures that are not only general-purpose but also lightweight and efficient.

To address these challenges, we introduce the *Multi-scale Temporal Network* (MSTN), a hybrid architecture grounded in the principle of *Early Temporal Aggregation* (ETA). The central idea of ETA is to decouple expressive temporal modeling from inference-time complexity by collapsing the temporal dimension early in the network via an $L \to 1$ transformation, where L denotes the lookback-window length. By performing the computationally intensive operations (e.g., $\mathcal{O}(L)$ BiLSTM or $\mathcal{O}(L^2)$ Transformer self-attention) *before* this aggregation step, the subsequent refinement layers—including feature fusion, SE recalibration, single dense layer (SDL), and prediction modules—operate with a fixed $\mathcal{O}(1)$ cost with respect to L. MSTN achieves this through a dual-path encoder that integrates complementary temporal strategies. Multi-scale convolutional encoders capture fine-grained local patterns, while a sequence modeling backbone (BiLSTM or Transformer) captures long-range dependencies. Their outputs are fused through a self-gated fusion (SGF) mechanism, augmented with squeeze-and-excitation (SE) recalibration and SDL, producing compact, discriminative, and temporally coherent representations.

We evaluate MSTN across forecasting, imputation, classification, and cross-domain generalization tasks, demonstrating consistent SOTA performance on 33 of 40 benchmark datasets. MSTN establishes new best results in long-term forecasting, achieves strong robustness under high missingness in imputation, and delivers superior classification accuracy across diverse application domains. In particular, these gains are achieved alongside efficiency improvements: the MSTN-BiLSTM and MSTN-Transformer variants utilize fixed cores of $\sim$278,520 parameters and $\sim$950,776 $\approx$ 1M parameters, respectively. The proposed work thus has a twofold impact. First, MSTN provides fast inference and can be embedded in real-time applications for time-series classification/forecasting, such as healthcare monitoring, industrial machine health, flight navigation, etc. Second, MSTN has a very small memory footprint and is thus suited for edge-device applications.

The key contributions of this work are summarized as follows:

1. We introduce MSTN, a multi-scale neural framework that is both fast and efficient. By leveraging the ETA principle to confine expensive $O(L^2)$ operations to the initial encoder, MSTN achieves a constant $\boldsymbol{O(1)}$ complexity with respect to the sequence length. MSTN incurs $\sim$**278,520** (MSTN-

BiLSTM) and ∼**950,776** (MSTN-Transformer) parameters and **sub-millisecond inference** latency.

2. Through an extensive evaluation across 40 standard benchmarks spanning imputation, long term forecasting, short term forecasting, classification, and cross-dataset generalization, MSTN achieves **SOTA performance on 33 datasets,** outperforming recent models such as DARTS-TS, LLM4TS, SOFTS, SimpleTM, TimeMixer, iTransformer, TimesNet, and PatchTST.

3. We also show the generalizability of MSTN on **7 datasets** from various domains such as health, safety, etc.

2 Related Work

In multivariate time series, a lot of work has been done. However, we are going to discuss only the most recent work in the literature of multivariate time series. Our discussion focuses on two aspects of time-series models: (i) the tasks they handle and (ii) efficiency.

2.1 Task Perspective

Work in time-series prediction can be categorized based on the following tasks: Imputation, long-term forecasting, short term forecasting, classification, and anomaly detection. In this section, we focus on the first four tasks.

The majority of works focus on long-term forecasting tasks, including specialized architectures — such as DLinear (Zeng et al., 2022), PatchTST (Nie et al., 2023),iTransformer (Liu et al., 2024), SOFTS (Han et al., 2024a), TIME-LLM (Jin et al., 2024), LLM4TS (Chang et al., 2025)), and masked autoencoders (HiMTM (Zhao et al., 2024)). Seg-MoE (Ortigossa & Segal, 2026) improves scalability and long-term dependency modeling via segment-wise Mixture-of-Experts routing, but is primarily designed for forecasting and does not explicitly handle irregular or event-driven patterns. vLinear (Yue et al., 2026) introduces an efficient linear multivariate forecaster, achieving strong performance; however, its linear formulation limits its capacity to capture complex nonlinear temporal dynamics. These models have driven significant performance gains, but have not been applied to other time-series tasks like imputation or classification, which we address in the present work. In addition, the aforementioned works claim to handle channel dependency (CI, CD, or CP) differently. CI works like DLinear (Zeng et al., 2022) and PatchTST (Nie et al., 2023) handle time-series where channels are independent and therefore cannot model cross-channel effects. CD models such as iTransformer (Liu et al., 2024), graph-based approaches such as TPGNN (Liu et al., 2022b) capture full cross-channel dependencies but scale poorly with channel count. CP methods like MCformer (Han et al., 2024b) and ModernTCN (donghao & wang xue, 2024) project the channel dimension into a latent space before interaction, trading expressivity for efficiency. While these methods show impressive performance, they are limited in capturing the multi-scale temporal dynamics within and across channels—a capability essential for modeling long-sequence dependencies. To combine the strengths of both paradigms, our work takes an intermediate route where we leverage both CI and CD.

Another important forecasting task is short-term forecasting, which focuses on predicting immediate future values over short horizons (minutes to hours), which is critical for real-time applications such as traffic management, autonomous navigation, and incident detection. Recent state-of-the-art approaches include TimeMixer (Wang et al., 2024), which uses multi-scale mixing to extract both fine-grained and coarse-grained features, and SCINet (Liu et al., 2022a), which employs recursive downsampling and interaction to capture multi-scale temporal patterns. PatchTST (Nie et al., 2023) applies patch-based tokenization to capture local patterns, while TimesNet (Wu et al., 2023) transforms 1D sequences into 2D representations to capture complex temporal variations. DLinear (Zeng et al., 2022) offers a simple but effective linear baseline, and FEDformer (Zhou et al., 2022) leverages frequency-domain attention for temporal modeling. Despite their strengths, these methods often rely on fixed-scale aggregations or require careful tuning of downsampling rates, limiting their adaptability to varying traffic dynamics. In contrast, MSTN's dual-branch design captures both rapid fluctuations and periodic patterns.

Another line of work in time-series prediction is imputation, which addresses the challenge of reconstructing missing values caused by sensor failures or data corruption, an essential step to maintain data integrity in real-world systems (Zhou et al., 2023; Wu et al., 2023). Imputation tasks employ specialized architectures such as TimesNet (Wu et al., 2023), GPT2(3) (Zhou et al., 2023), PatchTST (Nie et al., 2023), LightTS (Zhang et al., 2022), and DLinear (Zeng et al., 2022). However, these methods perform poorly on aperiodic or irregularly sampled events (e.g., emergency braking) because they have not been designed to detect such chaotic events. On the other hand, in the present work, MSTN can handle such events gracefully.

Another time series task, classification, involves categorizing the entire sequence to identify underlying events or patterns, enabling automated diagnosis and monitoring in fields such as healthcare and industrial maintenance (Wu et al., 2023; Zhang et al., 2022). Classification specialized architectures such as GPT2(6) (Alharthi et al., 2025), TimesNet (Wu et al., 2023), DLinear (Zeng et al., 2022), ETSFormer (Woo et al., 2022), FEDformer (Zhou et al., 2022), and Informer (Zhou et al., 2021). However, by design, these methods focus on seasonal decomposition, which can lead to over smoothing of temporal dynamics. Consequently, they often miss sudden, aperiodic anomalies such as spikes or crashes. They also tend to specialize in either local or global trends, rarely balancing both effectively.

2.2 Efficiency Perspective

In this section, we discuss efficiency of the time series models in terms of (i) characteristics of time series captured such as periodicity, varying temporal patterns, non-stationary behaviors, etc.; (ii) inference latency: and (iii) memory.

Periodic models such as TimesNet (Wu et al., 2023), Autoformer (Wu et al., 2022), ETSFormer (Woo et al., 2022), FEDformer (Zhou et al., 2022), LightTS (Zhang et al., 2022), PatchTST (Nie et al., 2023), and DLinear (Zeng et al., 2022) explicitly model periodic or seasonal structures and have demonstrated strong performance on long-term forecasting tasks. However, they often struggle to capture aperiodic, irregular, sudden spikes or event-driven dynamics. Dilated convolutional networks (e.g., ModernTCN (donghao & wang xue, 2024)) extract multiscale local patterns but have limited receptive fields for very long dependencies. Graph-based methods such as GTS (Shang et al., 2021) and FourierGNN (Yi et al., 2023) model spatio-temporal relationships but often assume static graphs, struggling with dynamically changing sensor relationships. Efficient attention variants (Informer (Zhou et al., 2021), FEDformer (Zhou et al., 2022), Autoformer (Wu et al., 2022)) use sparse or frequency-domain attention for long-range modeling. Despite these advances, these methods maintain the full sequence length throughout the network, incurring substantial memory and computational cost as the length of the sequence increases. They also tend to specialize in either local or global trends, rarely balancing both effectively while meeting the requirement for low-latency inference.

Recent MLP-based architectures such as TSMixer (Chen et al., 2023) and TTM (Ekambaram et al., 2024) achieve computational efficiency through temporal and channel mixing, demonstrating competitive performance with lower computational costs compared to attention-based models. However, these architectures employ fixed mixing operations that may not adapt dynamically to varying temporal patterns across different scales and frequencies. A distinct line of work employs large language models (LLM) for time series (e.g., TIME-LLM (Jin et al., 2024), LLM4TS (Chang et al., 2025)), leveraging their few-shot capabilities for cold-start scenarios. However, these models suffer from high latency and memory footprints, making them unsuitable for real-time edge deployment. Other methods, such as SOFTS (Han et al., 2024a), use stochastic pooling to reduce resolution but do not fully decouple sequence length from downstream processing, incur high inference latency and parameter counts. There is a need for architectures that are not only general-purpose but also lightweight and efficient. Our work attempts to address this gap.

3 Methodology

This section details the MSTN architecture and its methodological innovations, followed by the baseline models, and training configurations.

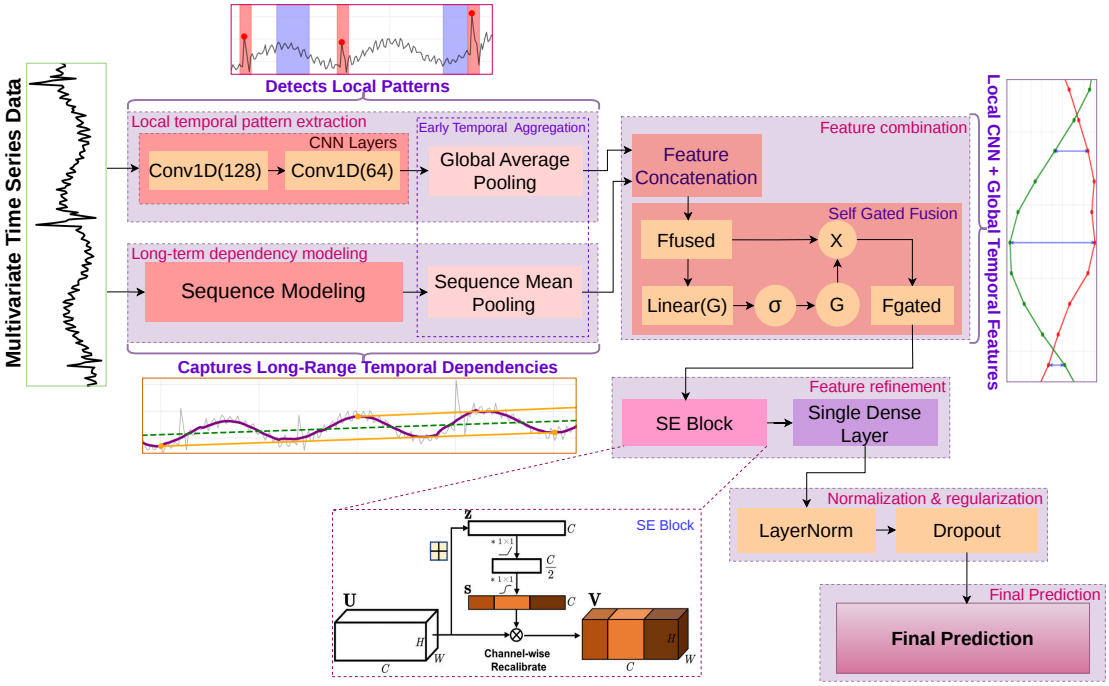

Figure 1: Architecture of the proposed Multi-scale Temporal Network (MSTN).

3.1 Proposed Architecture: MSTN

3.1.1 Architecture Overview

The proposed *Multi-scale Temporal Network* (MSTN; Fig. 1) is a hybrid deep-learning architecture designed around a parallel multi-scale modeling paradigm. MSTN processes temporal signals through two coordinated encoding branches, each specialized for complementary aspects of the temporal structure. A key design element is the use of an *early temporal aggregation* (ETA) mechanism, applied directly to the output of the dual encoders, which collapses the full input sequence of length L into a single fixed-dimensional representation ($L \to 1$). This operation ensures that all subsequent refinement stages operate in constant time with respect to sequence length, thereby enabling efficient processing of long temporal input.

The network comprises three principal modules. (i) A *multi-scale convolutional branch* captures localized temporal patterns using hierarchically stacked convolutions; its output is aggregated via global average pooling, promoting robustness and translation invariance. (ii) A *sequence modeling branch* captures long-range dependencies, instantiated using either a bidirectional LSTM (MSTN-BiLSTM) or a Transformer encoder (MSTN-Transformer). To maintain computational efficiency, this branch is compressed through sequence mean pooling, yielding a global summary vector. (iii) A *multi-scale fusion and refinement module* integrates the pooled representations from the two branches. The features are concatenated and passed through a SGF mechanism, which adaptively modulates the contributions of each branch. The fused representation is subsequently refined using channel-wise recalibration via SE blocks and is followed by a SDL, enabling the network to emphasize the most informative feature dimensions. The final representation is normalized, regularized through dropout, and projected to the task-specific output space. For classification tasks, we employ the standard cross-entropy loss defined in Eq. 14. Figure 2 provides an overview of the end-to-end processing pipeline, from parallel feature extraction to multi-scale fusion and refinement.

3.1.2 Design Rationale

The MSTN architecture is motivated by specific limitations in existing time series models. Transformer-based approaches excel at global dependency modeling but tend to oversmooth local fluctuations (Nguyen

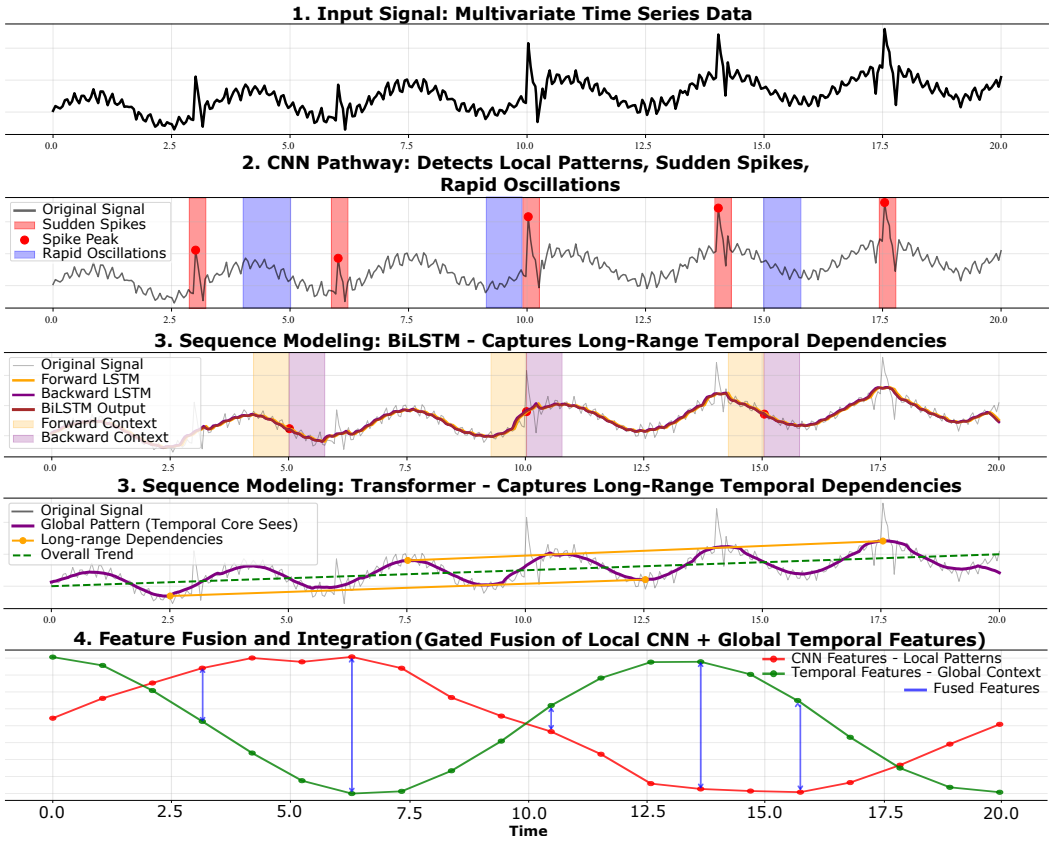

Figure 2: Multi-scale signal processing pipeline of MSTN.

et al., 2023; Noguchi & Kawahara, 2025), while convolutional networks capture fine-grained patterns but have limited receptive fields. Real-world time series are inherently multi-scale, with micro-fluctuations, sudden change events, and long-term trends coexisting—yet most architectures capture only one scale.

MSTN addresses this through a dual-branch design: a multi-scale CNN branch preserves local patterns that attention mechanisms would smooth, while a Transformer/BiLSTM branch models long-range dependencies beyond CNN reach. These complementary representations are dynamically balanced via a SGF mechanism that adaptively weights each branch based on input characteristics, rather than using fixed weights. SE blocks then recalibrate channel-wise feature importance, followed by a SDL that projects the fused representation into a higher-level feature space, enabling nonlinear combinations of the multi-scale temporal features while maintaining computational efficiency. Crucially, ETA collapses the temporal dimension ($L \to 1$) immediately after encoding, confining all $\mathcal{O}(L)$ and $\mathcal{O}(L^2)$ operations to the initial encoders and ensuring $\mathcal{O}(1)$ complexity for all subsequent processing. This combination yields an architecture that is simultaneously expressive (multi-scale), efficient (constant-time inference), and adaptable (dynamic fusion of complementary features), directly addressing the core limitations identified in existing approaches.

3.1.3 Mathematical Modeling

We consider the following problem: Given multivariate time series samples $X_{1:L} \in \mathbb{R}^{L \times C}$ with a lookback window sequence length L and C variate features, MSTN learns temporal representations through a novel parallel architecture. We detail the complete process below.

Forward Process: The input $X \in \mathbb{R}^{B \times L \times C}$ is processed through dual complementary branches that operate simultaneously. The CNN captures local temporal patterns, while the sequence modeling (implemented as either a Transformer or a BiLSTM) captures long-range temporal dependencies. Their outputs are fused

through a sophisticated gating and attention mechanism for enhanced representation learning. The input undergoes successive Conv1D operations to extract multi-scale temporal features, progressively adjusting the feature dimensionality while preserving temporal resolution. The Conv1D_k convolutional layers employ a hierarchical design in which the first layer with kernel size $k = 7$ extracts basic temporal patterns, and the second layer with kernel size $k = 5$ learns combinations of these patterns into more complex features, with padding to maintain temporal dimensions, and batch normalization stabilizes training. To aggregate these temporal features into a compact representation, we apply global average pooling across the entire sequence, where $X^\top \in \mathbb{R}^{B\times C\times L}$ denotes the channel-first representation obtained by transposing the input along the temporal and variable dimensions.

$$H^{(1)}_{\text{conv}} = \text{ReLU}(\text{BN}(\text{Conv1D}_7(X^T))) \in \mathbb{R}^{B\times 128\times L} \tag{1}$$

$$H^{(2)}_{\text{conv}} = \text{ReLU}(\text{BN}(\text{Conv1D}_5(H^{(1)}_{\text{conv}}))) \in \mathbb{R}^{B\times 64\times L} \tag{2}$$

$$\mathbf{z}_{\text{cnn}} = \text{GlobalAvgPool1D}(H^{(2)}_{\text{conv}}) \in \mathbb{R}^{B\times 64} \tag{3}$$

where each of the 64 feature channels is summarized by its average activation over time through temporal pooling. We employ two alternative sequence modeling approaches: MSTN-Transformer and MSTN-BiLSTM. First, the input sequence is projected into a 128-dimensional latent feature space using a linear embedding layer. Specifically, the projected representation is obtained as $X_p = W_p X + b_p$, where $X_p \in \mathbb{R}^{B\times L\times 128}$, and $W_p \in \mathbb{R}^{C\times 128}$ and $b_p \in \mathbb{R}^{128}$ denote the learnable weight matrix and bias of the projection layer that embed the original C-dimensional input features into a 128-dimensional latent space. In MSTN-Transformer, the architecture captures long-range dependencies through a *single* Transformer encoder composed of 4 layers with 8 attention heads. The encoder produces contextualized representations for each of the L time steps. To consolidate these temporal representations into a fixed-length global descriptor, we apply sequence mean pooling over the temporal dimension:

$$H_{\text{trans}} = \text{TransformerEnc}(X_p) \in \mathbb{R}^{B\times L\times 128}, \quad \mathbf{z}_{\text{trans}} = \frac{1}{L}\sum_{t=1}^{L} H_{\text{trans},t} \in \mathbb{R}^{B\times 128} \tag{4}$$

In MSTN-BiLSTM, bidirectional sequential dependencies are modeled by processing the projected sequence in both forward and backward directions, enabling the network to learn complex temporal dynamics and long-range contextual patterns:

$$H_{\text{bilstm}} = \text{BiLSTM}(X_p) \in \mathbb{R}^{B\times L\times 128}, \quad \mathbf{z}_{\text{bilstm}} = \frac{1}{L}\sum_{t=1}^{L} H_{\text{bilstm},t} \in \mathbb{R}^{B\times 128} \tag{5}$$

The BiLSTM consists of 2 layers with 64 hidden units per direction, capturing both causal and anti-causal temporal relationships.

The CNN branch utilizes global average pooling, and the sequence modeling branches employ sequence mean pooling. These pooling operations, along with the subsequent reshaping, collectively form the ETA module, which collapses the temporal dimension L to 1. This critical step ensures that all subsequent operations execute at $\mathcal{O}(1)$ inference cost. This design enables MSTN to achieve constant-time inference with respect to the input sequence length while preserving both local and global temporal semantics. The parallel features undergo sophisticated fusion through a multi-stage enhancement process that combines the strengths of both branches. For each MSTN variant (MSTN-Transformer or MSTN-BiLSTM):

$$\mathbf{z}_{\text{concat}} = [\mathbf{z}_{\text{cnn}}; \mathbf{z}_{\text{trans/bilstm}}] \in \mathbb{R}^{B\times 192} \tag{6}$$

where $\mathbf{z}_{\text{cnn}} \in \mathbb{R}^{B\times 64}$ and $\mathbf{z}_{\text{trans/bilstm}} \in \mathbb{R}^{B\times 128}$. The architecture then performs adaptive weighting through an SGF mechanism, illustrated in Figure 1, that learns to dynamically balance feature contributions:

$$\mathbf{z}_{\text{fused}} = \mathbf{z}_{\text{concat}} \odot \sigma(W_g \mathbf{z}_{\text{concat}} + b_g) \tag{7}$$

where for both variants: $W_g \in \mathbb{R}^{192\times 192}$, $b_g \in \mathbb{R}^{192}$. This mechanism learns to emphasize the most relevant features from each branch, creating a harmonious blend of multi-scale temporal features. The fused features are reshaped for efficient single-token processing: $\mathbf{z}_{\text{seq}} = \mathbf{z}_{\text{fused}} \otimes \mathbf{1} \in \mathbb{R}^{B\times 1\times 192}$. Channel-wise attention then enhances features through an SE block. Given that the temporal dimension is already $L = 1$, the global pooling step within SE is trivially executed: $\mathbf{z}_{\text{se}} = \mathbf{z}_{\text{seq}} \odot \sigma(W_2 \text{ReLU}(W_1 \mathbf{z}_{\text{seq}}))$ where $W_1 \in \mathbb{R}^{24\times 192}$, $W_2 \in \mathbb{R}^{192\times 24}$ with a reduction ratio 8. The SE mechanism amplifies informative channels while suppressing less useful ones. Feature dependencies are refined through a SDL. Since the temporal dimension is condensed ($L = 1$), the SDL functions as a sophisticated global feature recalibration layer rather than a temporal mixer, confirming $\mathcal{O}(1)$ complexity:

$$\mathbf{z}_{\text{final}} = \text{Dropout}\left(\text{LayerNorm}\left(W_d \cdot \mathbf{z}_{\text{se}} + b_d\right)\right) \tag{8}$$

where $W_d \in \mathbb{R}^{192\times 192}$ and $b_d \in \mathbb{R}^{192}$ are the learnable weight matrix and bias of the SDL respectively. Dropout with rate $p = 0.1$ provides regularization, and layer normalization ensures stable training.

The refined features $\mathbf{z}_{\text{final}}$ are then passed to the final task-specific head (imputation, forecasting, or classification). The prediction head consists of a single linear layer that maps the 192-dimensional aggregated feature vector to the task-specific output shape.

Task-Specific Prediction and Loss Functions: We detail the prediction head ($\hat{Y}$ or $\hat{X}$) and the corresponding loss function $\mathcal{L}$ for each of the three tasks. The final representation $\mathbf{z}_{\text{final}}$ is passed through specialized linear heads to map the latent features to the required output space.

Time Series Imputation: The head reconstructs the input sequence $\hat{X}_{1:L} \in \mathbb{R}^{B\times L\times C}$. The objective is to minimize the masked Mean Squared Error (MSE) computed only over masked positions, where Ω is the binary mask matrix ($\Omega_{i,t,c} = 1$ if the value is masked during training/evaluation):

$$\hat{X}_{1:L} = W_i \mathbf{z}_{\text{final}} + b_i \in \mathbb{R}^{B\times L\times C} \tag{9}$$

$$\mathcal{L}_{\text{imputation}} = \frac{1}{\|\Omega\|_1} \sum_{i=1}^{B} \sum_{t=1}^{L} \sum_{c=1}^{C} \Omega_{i,t,c} \cdot \left(\hat{X}_{i,t,c} - X_{i,t,c}\right)^2 \tag{10}$$

Long-Term Forecasting: The head produces a forecast $\hat{Y}_{1:H} \in \mathbb{R}^{B\times H\times C}$ for a prediction horizon H. The objective is to minimize the MSE across the batch, horizon, and variable dimensions:

$$\hat{Y}_{1:H} = W_f \mathbf{z}_{\text{final}} + b_f \in \mathbb{R}^{B\times H\times C} \tag{11}$$

$$\mathcal{L}_{\text{forecast}} = \frac{1}{BHC} \sum_{i=1}^{B} \sum_{t=1}^{H} \sum_{c=1}^{C} \left(\hat{Y}_{i,t,c} - Y_{i,t,c}\right)^2 \tag{12}$$

Although mapping a 192-dimensional latent vector to horizon H (especially $H = 720$) is theoretically compressed, our empirical results demonstrate improved performance. For example, our model achieves 0.181 MSE on ETTh1 at $H = 720$ and 0.151 MSE on ETTm1 at $H = 720$ (ranked 1st among baselines, Table 5). The linear decoder learns 192 basis temporal patterns, and the forecast is a linear combination $\hat{Y} = \sum_{i=1}^{192} \alpha_i \phi_i(t)$, where $\phi_i(t)$ are rows of W_f. This basis learning approach allows the decoder to represent complex dynamics—from low-frequency trends to high-frequency fluctuations—as weighted combinations of learned patterns, with coefficients α_i derived from the rich latent representation $\mathbf{z}_{\text{final}}$.

Table 1: Taxonomy of model architectures by Channel-Interaction strategy. Different state-of-the-art models align with the established Channel Interaction (CI, CD, CP) strategies. MSTN is positioned as a hybrid design (CI+CD) via early temporal aggregation to capture multi-scale dynamics. (Asym.: Asymmetry, Lag.: Lagginess, Pol.: Polarity, Gw.: group-wise, Dyn.: Dynamism, Ms.: Multi-scale)

Strategy	**Mechanism**	**Characteristic**						**Method**
		Asym.	**Lag.**	**Pol.**	**Gw.**	**Dyn.**	**Ms.**	
CI	-	-	-	-	-	-	-	PatchTST
CI	-	-	-	-	-	-	-	CycleNet
CI	-	-	-	-	-	-	-	DLinear
CI	-	-	-	-	-	-	-	Timer
CI	-	-	-	-	-	-	-	Chronos
CI	-	-	-	-	-	-	-	LLM4TS
CI	-	-	-	-	-	-	-	Time-LLM
CI	-	-	-	-	-	-	-	RevIN
CD	CNN-based	✓	-	-	-	-	-	Informer
CD	CNN-based	✓	-	-	-	-	-	Autoformer
CD	CNN-based	✓	-	-	-	-	-	FEDformer
CD	CNN-based	✓	-	-	-	-	-	TimesNet
CD	MLP-based	✓	-	-	-	-	-	TSMixer
CD	MLP-based	✓	-	-	-	-	-	TTM
CD	Transformer-based	✓	-	-	-	✓	-	iTransformer
CD	Transformer-based	✓	-	-	-	✓	-	Crossformer
CD	Transformer-based	✓	✓	-	-	-	-	VCformer
CD	Transformer-based	✓	✓	-	-	✓	-	MOIRAI
CD	Transformer-based	✓	-	-	-	✓	-	UniTS
CD	GNN-based	-	-	-	✓	-	-	GTS
CD	GNN-based	✓	-	-	-	-	✓	MSGNet
CD	GNN-based	-	✓	-	-	-	-	FourierGNN
CD	GNN-based	-	✓	-	-	-	-	FC-STGNN
CD	GNN-based	✓	-	-	-	✓	-	TPGNN
CD	Linear Channel Mixing (Factorization)	✓	-	-	-	-	-	SOFTS
CD	Others	✓	-	-	-	✓	-	C-LoRA
CP	CNN-based	✓	-	-	-	-	-	ModernTCN
CP	Transformer-based	✓	-	-	✓	-	-	DUET
CP	Transformer-based	✓	-	-	-	✓	-	MCformer
CP	Transformer-based	✓	-	-	✓	✓	-	DGCformer
CP	Transformer-based	-	-	-	-	✓	-	CM
CP	GNN-based	✓	-	-	-	-	-	MTGNN
CP	GNN-based	✓	-	✓	-	-	-	CrossGNN
CP	GNN-based	✓	-	-	✓	-	-	WaveForM
CP	GNN-based	-	-	-	-	✓	-	MTSF-DG
CP	GNN-based	✓	-	-	✓	-	-	ReMo
CP	GNN-based	✓	-	-	✓	✓	✓	Ada-MSHyper
CP	Others	✓	✓	-	-	-	-	LIFT
CP	Others	✓	-	-	✓	-	-	CCM
CI + CD	**Parallel CNN (CI) and BiLSTM/Transformer(CD) based Dual Encoder + ETA**	✓	-	-	-	✓	✓	**MSTN**

Time Series Classification: The head computes class probabilities $P \in \mathbb{R}^{B \times K}$ over K target classes. The model is optimized using the canonical Cross-Entropy (CE) Loss:

$$P = \text{Softmax}(W_c \mathbf{z}_{\text{final}} + b_c) \in \mathbb{R}^{B \times K} \tag{13}$$

$$\mathcal{L}_{\text{classify}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{K} Y_{i,k} \log(P_{i,k}) \tag{14}$$

where Y is the one-hot encoded true label. This architectural modularity allows the MSTN framework to be universally applied to diverse multivariate time series tasks without structural modification.

3.2 How is MSTN different from prior works?

The established field of multivariate time–series forecasting is conventionally organized around three channel interaction strategies—CI, CD, and CP approaches (Qiu et al., 2025)—as summarized in Table 1. These strategies specify how a model encodes and exploits inter-variable dependencies across the C channels of a multivariate signal.

Table 2: Comparison of architectural characteristics. Evaluation of the MSTN framework against the CI, CD, and CP Channel Interaction strategies. MSTN achieves high capacity while neutralizing the $\mathcal{O}(L^2)$ computational bottleneck.

Dimension	CI	CD	CP	(CI+CD) MSTN	Rationale
Efficiency	High	Low	Moderate	High	ETA enables $\mathcal{O}(1)$ refinement after initial encoding.
Robustness	High	Low	Moderate	High	Dual-encoder with SGF provides complementary feature stability.
Generalizability	Low	Moderate	High	High	Achieved via a single, aggregated feature vector ($\mathbf{z}_{\text{final}}$) that supports multi-task learning (Forecasting, Imputation, Classification).
Capacity	Low	High	Moderate	High	Transformer/BiLSTM path captures global dependencies.
Ease of Implementation	High	Moderate	Low	Moderate	Relies on standard components (CNN, Transformer/BiLSTM, SGF, SE, SDL) rather than complex, specialized graph or recursive structures.

CI. CI methods (e.g., DLinear, PatchTST) process each variable independently using lightweight temporal operators such as linear layers or multilayer perceptrons. Their advantage lies in linear channel cost $\mathcal{O}(C)$ and efficient temporal complexity $\mathcal{O}(L)$, though this comes at the expense of discarding cross-channel structure.

CD. CD approaches (e.g., iTransformer) explicitly model inter-variable correlation through dense attention or graph-based mechanisms. While these models offer comprehensive feature coupling, they incur high computational cost—typically $\mathcal{O}(C^2)$ or $\mathcal{O}(L^2)$—which limits their scalability and latency characteristics.

CP. CP models (e.g., MCformer, MTGNN) map the channel dimension into a low-rank latent space before applying interaction, thereby avoiding the full $\mathcal{O}(C^2)$ cost of CD models. Although channel-efficient, these models often impose simplified temporal assumptions, reducing their ability to capture multi-scale dynamics.

By combining the complementary strengths of CI and CD models with a novel temporal aggregation mechanism, MSTN delivers strong performance across forecasting, imputation, and classification tasks while maintaining the low-latency characteristics.

3.2.1 Core Innovation: Dual-Path Design and ETA

The MSTN framework integrates a dual-path temporal encoding architecture with the ETA mechanism, resulting in a principled hybrid approach that reconciles the core trade-offs among CI-, CD-, and CP-based modeling strategies as shown in Table 2. Its design is centered on two architectural innovations that jointly enhance representational fidelity and computational efficiency:

1. **Dual-branch temporal encoding.** In contrast to recent CP-oriented architectures such as MCformer (Han et al., 2024b), which priorities efficiency at the expense of temporal expressiveness, MSTN explicitly preserves multi-scale temporal structure. It deploys a parallel encoding strategy in which a global sequence-modeling branch (Transformer or BiLSTM; CD strategy) captures long-range temporal dependencies, while a lightweight convolutional branch (CI strategy; $\mathcal{O}(L)$) extracts fine-grained local temporal patterns. This duality directly addresses the multi-scale nature of real-world temporal signals, offering coverage beyond what is typically afforded by efficient CP-based methods.

2. **Efficiency through ETA: collapsing $\mathcal{O}(L^2)$ to $\mathcal{O}(1)$.** MSTN applies the ETA mechanism immediately after the high-capacity encoders, collapsing the temporal dimension ($L \to 1$) via learned sequence aggregation. By performing the computationally intensive operations (e.g., $\mathcal{O}(L^2)$ Transformer self-attention) *before* this aggregation step, the subsequent refinement layers—including feature fusion, SE recalibration, SDL, and prediction modules—operate with a fixed $\mathcal{O}(1)$ cost with respect to L. This explicitly reshapes the model's complexity profile, enabling MSTN to maintain the representational power of CD architectures while achieving substantially improved inference efficiency. Unlike conventional temporal pooling, hierarchical downsampling, or token-merging strategies, ETA is not introduced as a representational shortcut, but as a principled reordering of computation. Prior approaches typically aggregate temporal information either progressively across

layers or at the final prediction stage, implicitly coupling representational capacity with inference-time complexity. In contrast, ETA explicitly allows high-capacity temporal modeling to operate on the full input sequence, while enforcing an early collapse of the temporal dimension before downstream fusion and refinement.

4 Experiments

This section presents a comprehensive evaluation of the proposed MSTN model. We demonstrate its performance on four time-series tasks: (1) imputation, (2) long-term forecasting, (3) short-term forecasting, (4) classification. Additionally, we also present the generalizability study on standard time-series benchmark datasets.

4.1 Baselines

This section presents the baselines used to assess MSTN on five time series tasks: (1) imputation, (2) long-term forecasting, (3) short-term forecasting (4) classification, and (5) generalizability study.

4.1.1 Imputation

Following the TimesNet (Wu et al., 2023) protocol, we evaluate performance under the Missing Completely at Random (MCAR) scenario with random masking ratios of $\{12.5\%, 25\%, 37.5\%, 50\%\}$. Performance is quantified using Mean Squared Error (MSE) and Mean Absolute Error (MAE). The evaluation includes recent SOTA baselines such as GPT2(3) (Zhou et al., 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), LightTS (Zhang et al., 2022), and DLinear (Zeng et al., 2022).

4.1.2 Long Term Forecasting

The proposed MSTN model is evaluated for the long-term time series forecasting task. We evaluated all methods across multiple prediction horizons $H \in \{96, 192, 336, 720\}$ for standard benchmarks and $H \in \{24, 36, 48, 60\}$ for the ILI dataset. The forecast performance is quantified using MSE and MAE. The evaluation includes the following SOTA baselines: DARTS-TS (Deng & Lindauer, 2025), LLM4TS (Chang et al., 2025), HiMTM (Zhao et al., 2024), TIME-LLM (Jin et al., 2024), MTST (Zhang et al., 2024), SOFTS (Han et al., 2024a) and iTransformer (Liu et al., 2024).

We also evaluated the proposed MSTN model for forecasting on four public subsets of California traffic network data from the Performance Measurement System (PEMS), namely PEMS03, PEMS04, PEMS07, and PEMS08, with prediction horizons $H \in \{12, 24, 36, 48\}$, following the protocol of DARTS-TS (Deng & Lindauer, 2025), and iTransformer (Liu et al., 2024). We compare our model against state-of-the-art baselines including DARTS-TS (Deng & Lindauer, 2025), iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2023), TSMixer (Chen et al., 2023) and TimesNet (Wu et al., 2023).

4.1.3 Short-Term Forecasting

Following the benchmark protocol established by SimpleTM (Chen et al., 2025), TimeMixer (Wang et al., 2024) and SCINet (Liu et al., 2022a), we evaluate MSTN on short-term forecasting tasks using PEMS traffic datasets (PEMS03, PEMS04, PEMS07, PEMS08) with a prediction horizon $H = 12$. Performance is quantified using MAE, Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) on the original scale. The evaluation includes recent state-of-the-art baselines such as SimpleTM (Chen et al., 2025), TimeMixer (Wang et al., 2024), SCINet (Liu et al., 2022a), PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023), DLinear (Zeng et al., 2022), and FEDformer (Zhou et al., 2022).

4.1.4 Classification

The performance of classification tasks is evaluated using the standard metric of classification accuracy. The evaluation compares with recent SOTA baselines such as GPT2(6) (Zhou et al., 2023), TimesNet (Wu et al.,

2023), DLinear (Zeng et al., 2022), ETSFormer (Woo et al., 2022), FEDformer (Zhou et al., 2022), and Informer (Zhou et al., 2021).

4.1.5 Generalizability Study

To evaluate the generalizability of the proposed MSTN framework across domains, we benchmarked it against TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), and prior work models from the respective literature (e.g., Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), k-means, Naive Bayes (NB), k-Nearest Neighbors (k-NN), and Stacked Autoencoder (SAE)). All evaluations utilized a consistent five-fold cross-validation framework on seven publicly available international datasets spanning healthcare, activity recognition, agricultural technology, and industrial monitoring.

4.2 Experimental Setup

4.2.1 Data Preprocessing and Evaluation Protocol

Our preprocessing and splits follow established practices from prior work (Zhou et al., 2023; Wu et al., 2023; Nie et al., 2023). All datasets are split chronologically to preserve temporal causality. For ETT datasets (ETTh1, ETTh2, ETTm1, ETTm2), we use a 60%/20%/20% train/validation/test split. For other datasets such as ECL, Exchange and Traffic, we adopt a 70%/10%/20% split for forecasting and imputation tasks. The raw time series is partitioned before constructing sliding windows to avoid temporal overlap across splits. All features are normalized using train-only statistics, where the mean and standard deviation are computed from the training set and applied to the validation and test sets. For forecasting and imputation tasks, evaluation metrics (MSE and MAE) are computed on normalized predictions consistent with established benchmarks (Nie et al., 2023; Wu et al., 2023; Liu et al., 2024). Furthermore, we evaluated the proposed model on four PEMS datasets (PEMS03, PEMS04, PEMS07, PEMS08) for forecasting, following the protocol of DARTS-TS (Deng & Lindauer, 2025), and iTransformer (Liu et al., 2024). We use a lookback window of $L = 96$ except for ILI where $L = 36$, and prediction horizons $H \in \{96, 192, 336, 720\}$ for standard datasets and $H \in \{24, 36, 48, 60\}$ for ILI. The data is split chronologically into 60% training, 20% validation, and 20% testing. Metrics (MSE and MAE) are reported on the normalized scale, consistent with prior benchmarks. For short-term forecasting, we evaluated four PEMS datasets (PEMS03, PEMS04, PEMS07, PEMS08) with $L = 96$, $H = 12$, and a chronological 60/20/20 split following the protocols of TimeMixer (Wang et al., 2024) and SCINet (Liu et al., 2022a). Metrics MAE, MAPE and RMSE are reported on the original scale after inverse transformation.

4.2.2 Training and Model Configuration

The proposed MSTN model is implemented in Python 3.13.1 using PyTorch 2.7.1 and trained on an NVIDIA DGX A100 server equipped with 8 $\times$ NVIDIA A100-SXM4 GPUs (40 GB HBM2e VRAM each), using CUDA 12.3 and driver version 535.54.03. The architecture processes variable-length sequences through either a sequence modeling core (BiLSTM hidden units: 64 per direction (128 total)) or a Transformer (4 layers, 8 attention heads), combined with a CNN branch (128 $\rightarrow$ 64 filters). Training is performed using the AdamW (Kingma & Ba, 2017) optimizer with a learning rate of 1×10^{-4}, batch size of 64, and task-specific loss functions. For forecasting and imputation tasks, Mean Squared Error (MSE) loss is used, while for classification tasks, Cross-Entropy Loss is employed. For imputation, a masked MSE loss is applied, where the loss is computed only over randomly masked time points. All experiments follow the preprocessing protocol described in Section 4.2.1, including train-only standardization (Z-score normalization). For forecasting and imputation tasks, a fixed input lookback window of $L = 96$ (or $L = 36$ for ILI) is used across all prediction horizons to ensure consistency with prior benchmarks. All models are trained for up to 50 epochs with early stopping based on validation performance. Hyperparameters are tuned following standard practice in time-series forecasting literature (Madhusudhanan et al., 2024; Fristiana et al., 2024). To ensure robustness and reproducibility, all experiments are repeated with 5 different random seeds and results are reported as mean $\pm$ standard deviation.

Table 3: Benchmark Datasets for Forecasting, Imputation and Classification (10 UEA)

Task	Datasets	Features	Timesteps	Series Length	Information (Frequency)
Imputation	ETTh1/ETTh2	7	17,420	96	Electricity (15 mins)
	ETTm1/ETTm2	7	69,680	96	Electricity (15 mins)
	Electricity	321	26,304	96	Electricity (15 mins)
	Weather	21	52,696	96	Weather (10 mins)
Forecasting	ETTh1/ETTh2	7	17,420	{96,192,336,720}	Electricity (15 mins)
	ETTm1/ETTm2	7	69,680	{96,192,336,720}	Electricity (15 mins)
	Electricity	321	26,304	{96,192,336,720}	Electricity (Hourly)
	Traffic	862	17,544	{96,192,336,720}	Transportation (Hourly)
	Weather	21	52,696	{96,192,336,720}	Weather (10 mins)
	Exchange-Rate	8	7,588	{96,192,336,720}	Exchange rate (Daily)
	ILI	7	966	{24,36,48,60}	Illness (Weekly)
	PEMS03	358	26,209	{12,24,36,48}	Transportation (5 mins)
	PEMS04	307	16,992	{12,24,36,48}	Transportation (5 mins)
	PEMS07	883	28,224	{12,24,36,48}	Transportation (5 mins)
	PEMS08	170	17,856	{12,24,36,48}	Transportation (5 mins)
Classification	EthanolConcentration	3	1,751	1751	Alcohol Industry
	FaceDetection	144	62	62	Face (250Hz)
	Handwriting	3	152	152	Handwriting
	Heartbeat	61	405	405	Heart Beat
	JapaneseVowels	12	29	29	Voice
	PEMS-SF	963	144	144	Transportation (Daily)
	SelfRegulationSCP1	6	896	896	Health (256Hz)
	SelfRegulationSCP2	7	1,152	1152	Health (256Hz)
	SpokenArabicDigits	13	93	93	Voice (11025Hz)
	UWaveGestureLibrary	3	315	315	Gesture

4.2.3 Evaluation Metrics

The performance of long-term forecasting (including PEMS with horizons $H \in \{12, 24, 48, 96\}$) and imputation tasks is evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE), computed over all prediction horizons and averaged across all variables. For short-term forecasting on PEMS datasets, we report MAE, Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) on the original scale, following the TimeMixer (Wang et al., 2024) and SCINet (Liu et al., 2022a) protocols. For classification tasks, we report Accuracy as the primary metric, following standard practice in the literature (e.g., GPT-2 (Zhou et al., 2023), TimesNet (Wu et al., 2023)).

4.3 Benchmark Datasets

We evaluate the proposed method across four fundamental time series tasks: imputation, long-term forecasting, short-term forecasting, and classification using established public benchmark datasets. Table 3 provides a detailed overview of datasets. Missing value imputation is performed under a MCAR setting, with masking ratios ranging from 12.5%, 25%, 37.5% and 50%. Benchmarks include ETTh1, ETTh2, ETTm1, ETTm2, Electricity, and Weather datasets (Zhou et al., 2021; Trindade, 2015; Köllé, 2025). Imputation results are quantified using MSE and MAE over sequence lengths of 96 time steps.

Long-term forecasting performance is assessed on nine widely-used benchmarks: ETTh1, ETTh2 (Electricity Transformer Temperature datasets of the China State Grid, recorded at an hourly resolution), ETTm1, ETTm2 (15-minute variants of the same transformer monitoring system), Electricity (ECL - hourly household consumption data from 321 U.S. clients), Traffic (hourly road occupancy rates from CalTrans loop detectors in the United States), Weather (10-minute meteorological observations from 21 U.S. stations), Exchange (daily currency exchange rates across major global economies), and ILI (weekly influenza-like illness records from the U.S. CDC). These datasets are extensively used for time-series benchmarking and are publicly available (Wu et al., 2022). Forecasting horizons are set between 96, 192, 336, and 720 time steps. Long-term forecasting performance is evaluated using MSE and MAE.

Additionally, we evaluated four PEMS traffic datasets (PEMS03, PEMS04, PEMS07, PEMS08) for both long-term forecasting and short-term forecasting, following the protocols of DARTS-TS (Deng & Lindauer, 2025), SCINet (Liu et al., 2022a) and TimeMixer (Wang et al., 2024). These datasets are extensively used for time-series benchmarking and are publicly available (Lin, 2025; Liu et al., 2022a; Chen & data, 2025;

Table 4: Imputation Task Results Comparison. We randomly mask 12.5%, 25%, 37.5%, and 50% time points to compare model performance under different missing degrees. Red/Blue: First/Second ranks. We evaluate each model five times and report the mean (standard deviation); results marked with underline are not significantly worse than the first rank (Wilcoxon signed-rank test with Holm–Bonferroni correction, $\alpha = 0.05$). Example: 0.010(0.001) means 0.010 ± 0.001. Tra.: Transformer, BiL.: BiLSTM, FED.: FEDformer.

Models	**MSTN-Tra. (Ours)**		**MSTN BiL. (Ours)**		**GPT2(3) (2023)**		**TimesNet (2023)**		**PatchTST (2023)**		**LightTS (2022)**		**DLinear (2023)**		**FED. (2022)**		**Stationary (2022)**	
Mask R.	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
ETTh1																		
12.5%	0.002(0.000)	0.013(0.000)	0.007	0.022	0.043	0.140	0.057	0.159	0.093	0.201	0.240	0.345	0.151	0.267	0.070	0.190	0.060	0.165
25%	0.004(0.001)	0.025(0.002)	0.012	0.042	0.054	0.156	0.069	0.178	0.107	0.217	0.265	0.364	0.180	0.292	0.106	0.236	0.080	0.189
37.5%	0.007(0.001)	0.038(0.002)	0.017	0.062	0.072	0.180	0.084	0.196	0.120	0.230	0.296	0.382	0.215	0.318	0.124	0.258	0.102	0.212
50%	0.011(0.001)	0.057(0.003)	0.022	0.081	0.107	0.216	0.102	0.215	0.141	0.248	0.334	0.404	0.257	0.347	0.165	0.299	0.133	0.240
Avg	0.006	0.033	0.015	0.052	0.069	0.173	0.078	0.187	0.115	0.224	0.284	0.373	0.201	0.306	0.117	0.246	0.094	0.201
ETTh2																		
12.5%	0.003(0.000)	0.016(0.001)	0.014	0.033	0.039	0.125	0.040	0.130	0.057	0.152	0.101	0.231	0.100	0.216	0.095	0.212	0.042	0.133
25%	0.006(0.000)	0.029(0.001)	0.028	0.065	0.044	0.135	0.046	0.141	0.061	0.158	0.115	0.246	0.127	0.247	0.137	0.258	0.049	0.147
37.5%	0.010(0.000)	0.046(0.001)	0.044	0.100	0.051	0.147	0.052	0.151	0.067	0.166	0.126	0.257	0.158	0.276	0.187	0.304	0.056	0.158
50%	0.017(0.001)	0.070(0.001)	0.062	0.138	0.059	0.158	0.060	0.162	0.073	0.174	0.136	0.268	0.183	0.299	0.232	0.341	0.065	0.170
Avg	0.009	0.040	0.037	0.084	0.048	0.141	0.049	0.146	0.065	0.163	0.119	0.250	0.142	0.259	0.163	0.279	0.053	0.152
ETTm1																		
12.5%	0.002(0.000)	0.009(0.000)	0.003	0.011	0.017	0.085	0.019	0.092	0.041	0.130	0.075	0.180	0.058	0.162	0.035	0.135	0.026	0.107
25%	0.004(0.001)	0.020(0.002)	0.005	0.021	0.022	0.096	0.023	0.101	0.044	0.135	0.093	0.206	0.080	0.193	0.052	0.166	0.032	0.119
37.5%	0.006(0.001)	0.031(0.002)	0.007	0.032	0.029	0.111	0.029	0.111	0.049	0.143	0.113	0.231	0.103	0.219	0.069	0.191	0.039	0.131
50%	0.009(0.002)	0.042(0.002)	0.011	0.044	0.040	0.128	0.036	0.124	0.055	0.151	0.134	0.255	0.132	0.248	0.089	0.218	0.047	0.145
Avg	0.005	0.026	0.007	0.027	0.028	0.105	0.027	0.107	0.047	0.140	0.104	0.218	0.093	0.206	0.062	0.177	0.036	0.126
ETTm2																		
12.5%	0.003(0.000)	0.013(0.000)	0.004	0.014	0.017	0.076	0.018	0.080	0.026	0.094	0.034	0.127	0.062	0.166	0.056	0.159	0.021	0.088
25%	0.006(0.001)	0.026(0.001)	0.007	0.028	0.020	0.080	0.020	0.085	0.028	0.099	0.042	0.143	0.085	0.196	0.080	0.195	0.024	0.096
37.5%	0.011(0.002)	0.041(0.002)	0.012	0.043	0.022	0.087	0.023	0.091	0.030	0.104	0.051	0.159	0.106	0.222	0.110	0.231	0.027	0.103
50%	0.016(0.002)	0.058(0.002)	0.019	0.061	0.025	0.095	0.026	0.098	0.034	0.110	0.059	0.174	0.131	0.247	0.156	0.276	0.030	0.108
Avg	0.009	0.035	0.011	0.037	0.021	0.084	0.022	0.088	0.029	0.102	0.046	0.151	0.096	0.208	0.101	0.215	0.026	0.099
ECL																		
12.5%	0.027(0.000)	0.042(0.000)	0.037	0.050	0.080	0.194	0.085	0.202	0.055	0.160	0.102	0.229	0.092	0.214	0.107	0.237	0.093	0.210
25%	0.054(0.000)	0.083(0.000)	0.065	0.094	0.087	0.203	0.089	0.206	0.065	0.175	0.121	0.252	0.118	0.247	0.120	0.251	0.097	0.214
37.5%	0.081(0.001)	0.120(0.001)	0.100	0.140	0.094	0.211	0.094	0.213	0.076	0.189	0.141	0.273	0.144	0.276	0.136	0.266	0.102	0.220
50%	0.110(0.001)	0.166(0.001)	0.146	0.201	0.101	0.220	0.100	0.221	0.091	0.208	0.160	0.293	0.175	0.305	0.158	0.284	0.108	0.228
Avg	0.068	0.103	0.087	0.121	0.090	0.207	0.092	0.210	0.072	0.183	0.131	0.262	0.132	0.260	0.130	0.259	0.100	0.218
Weather																		
12.5%	0.005(0.001)	0.014(0.001)	0.007	0.017	0.026	0.049	0.025	0.045	0.029	0.049	0.047	0.101	0.039	0.084	0.041	0.107	0.027	0.051
25%	0.012(0.001)	0.026(0.001)	0.015	0.031	0.028	0.052	0.029	0.052	0.031	0.053	0.052	0.111	0.048	0.103	0.064	0.163	0.029	0.056
37.5%	0.018(0.002)	0.037(0.002)	0.021	0.043	0.033	0.060	0.031	0.057	0.035	0.058	0.058	0.121	0.057	0.117	0.107	0.229	0.033	0.062
50%	0.024(0.002)	0.054(0.002)	0.030	0.062	0.037	0.065	0.034	0.062	0.038	0.063	0.065	0.133	0.066	0.134	0.183	0.312	0.037	0.068
Avg	0.015	0.033	0.018	0.038	0.031	0.056	0.030	0.054	0.033	0.056	0.055	0.117	0.052	0.110	0.099	0.203	0.032	0.059

Wu et al., 2022). Long-term forecasting performance is evaluated using MSE and MAE, while short-term forecasting on PEMS uses MAE, MAPE, and RMSE.

The multivariate time series classification is evaluated on ten diverse datasets from the University of East Anglia (UEA) archive (Bagnall et al., 2018), which includes applications from healthcare to activity recognition. To rigorously assess cross-domain transferability, we utilize seven publicly available international datasets spanning diverse applications: Human Safety/Healthcare (Fall Event (Boubezoul et al., 2020), Risky Driving (Rodegast et al., 2024a)); Human Activity Recognition (HAR) (UCI-HAR (Reyes-Ortiz et al., 2013), PAMAP2 (Reiss, 2012)); Agricultural Technology/Welfare (ActBe-Calf (Dissanayake et al., 2025)); and Industrial Monitoring/Predictive Maintenance (MetroPT-3 (Davari et al., 2021a), NASA Turbofan Engine (Saxena & Goebel, 2008)).

4.4 Imputation Results

Imputation is the task of filling in missing values in a dataset. In real-world time series data, sensors can fail or data can be corrupted, which may create gaps in the data. This task tests the ability of the model to intelligently reconstruct missing data based on the surrounding context, which is vital for maintaining data quality.

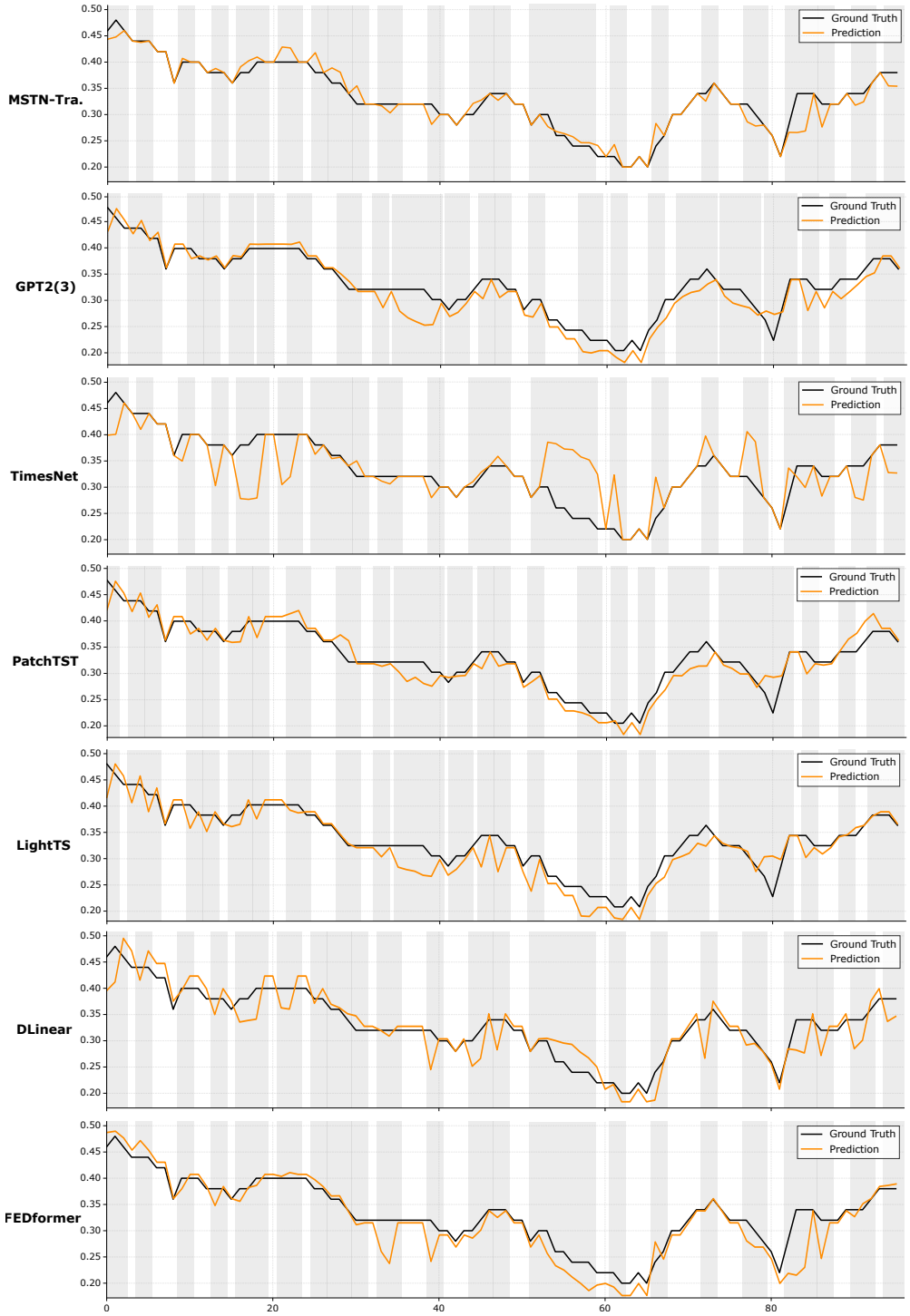

Figure 3: Visualization of weather imputation results under 50% mask ratio. The black lines represent ground truth and the orange lines represent predicted values.

Following the TimesNet (Wu et al., 2023) protocol, we evaluate imputation performance on electricity and weather domain datasets, including ETT (Zhou et al., 2021), ECL (Trindade, 2015), and Weather (Köllé, 2025). We evaluate performance under the MCAR scenario with random masking ratios of $\{12.5\%, 25\%, 37.5\%, 50\%\}$. During training, we randomly mask the specified percentage of time points. The

model reconstructs the complete sequence, but the loss is computed only on masked positions to ensure the model learns to impute missing values rather than memorize observed ones. At test time, we apply the same masking protocol and evaluate imputation performance by computing MSE and MAE exclusively on the masked positions. This ensures that the model is assessed solely on its imputation capability. Performance is quantified using MSE and MAE. The evaluation includes recent SOTA baselines such as GPT2(3) (Zhou et al., 2023), TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), LightTS (Zhang et al., 2022), and DLinear (Zeng et al., 2022).

As presented in Table 4, MSTN-Transformer achieves SOTA performance in 46 out of 48 imputation evaluation scenarios, delivering improved average results on all 6 benchmark datasets. For instance, on ETTh1 at 12.5% masking, MSTN-Transformer achieves MSE of 0.002, outperforming TimesNet (0.057) by 96.5%. On the challenging Weather dataset at 50% masking, our model achieves MSE of 0.024, surpassing GPT2(3) (0.037) by 35.1%. MSTN-BiLSTM demonstrates competitive performance, ranking second across all evaluation scenarios. These results validate the effectiveness of MSTN's multi-scale design for robust missing value reconstruction under varying missingness levels.

Figure 3 visualizes imputation results on real-world weather data with a 50% mask ratio across all evaluated methods—MSTN-Transformer, GPT2(3), TimesNet, PatchTST, LightTS, DLinear and FEDformer. The results show that MSTN-Transformer closely aligns with the ground truth temporal dynamics. Most baselines follow the general shape of the ground truth, but struggle with finer-grained temporal dynamics. In contrast, MSTN-Transformer consistently tracks both smooth trends and sharp fluctuations, indicating its effectiveness in capturing multi-scale temporal dependencies.

4.5 Long-Term Forecasting Results

Time series forecasting plays a crucial role in applications such as weather prediction, traffic management, and energy consumption planning. Long-term forecasting aims to predict future values far ahead in time. For example, it involves using historical weather data to estimate not only the temperature of tomorrow but also the conditions over the coming days or even weeks.

Multivariate long-term forecasting: We evaluate long-term forecasting performance using established protocols from the prior literature (Nie et al., 2023; Wu et al., 2023; 2022) across nine diverse benchmark datasets. The prediction horizons are set to $H \in \{96, 192, 336, 720\}$ for the standard benchmarks and $H \in \{24, 36, 48, 60\}$ for the ILI dataset. As presented in Table 5, MSTN-Transformer and MSTN-BiLSTM show improved performance compared to specialized time series forecasting architectures, including DARTS-TS (Deng & Lindauer, 2025), LLM4TS (Chang et al., 2025), HiMTM (Zhao et al., 2024), TIME-LLM (Jin et al., 2024), MTST (Zhang et al., 2024), SOFTS (Han et al., 2024a) and iTransformer (Liu et al., 2024). MSTN-Transformer and MSTN-BiLSTM demonstrate improved performance across multiple benchmark datasets. MSTN-Transformer achieves the best overall performance on ETTh1 (Avg MSE 0.138, MAE 0.286), ETTh2 (Avg MSE 0.184, MAE 0.341), and Weather (Avg MSE 0.205 and 0.307), while MSTN-BiLSTM leads on Exchange (Avg MSE 0.348 and MAE 0.404), and ILI dataset. On the Traffic dataset, HiMTM achieves the best average MSE 0.384 and MAE 0.254, while MSTN-Transformer delivers competitive performance with 0.578 MSE. Overall, our proposed architecture achieves improved results, demonstrating its effectiveness for long-term time series forecasting.

Long-Term Forecasting on PEMS dataset: Following the protocol of DARTS-TS (Deng & Lindauer, 2025) and iTransformer (Liu et al., 2024), we evaluate MSTN on four PEMS traffic datasets (PEMS03, PEMS04, PEMS07, PEMS08) with lookback window $L = 96$ and prediction horizons $H \in \{12, 24, 48, 96\}$. Table 6 presents the results with MSE and MAE reported on the **normalized scale**. MSTN-Transformer and MSTN-BiLSTM show improved performance compared to baselines including DARTS-TS (Deng & Lindauer, 2025), iTransformer (Liu et al., 2024), PatchTST (Nie et al., 2023), TSMixer (Chen et al., 2023) and TimesNet (Wu et al., 2023). On PEMS03 at $H = 12$, MSTN-Transformer achieves MSE of 0.010, outperforming DARTS-TS (0.066) by 83.3% and TimesNet (0.085) by 87.1%. On PEMS07, our model achieves an average MSE of 0.010, outperforming ModernTCN (0.304) by 96.7% and TSMixer (0.150) by 93.3%. On PEMS08, MSTN-Transformer achieves an average MSE of 0.013, surpassing DARTS-TS (0.169) by 92.3% and TimesNet (0.191) by 93.2%. These results demonstrate that MSTN's multi-scale design effectively cap-

Table 5: Multivariate long-term forecasting results. Prediction horizons $H \in \{96, 192, 336, 720\}$ ($H \in \{24, 36, 48, 60\}$ for ILI). Avg is averaged from all four prediction lengths. Red/Blue: First/Second ranks. We evaluate each model five times and report the mean (standard deviation); results marked with underline are not significantly worse than the first rank (Wilcoxon signed-rank test with Holm–Bonferroni correction, $\alpha = 0.05$). iTransf: iTransformer.

Models	MSTN-Tra. Ours		MSTN BiL. Ours		DARTS-TS 2025		LLM4TS 2025		HiMTM 2025		TIME-LLM 2024		MTST 2024		SOFTS 2024		iTransf. 2024	
H-Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1																		
96	0.076 (0.00)	0.216 (0.00)	0.101	0.240	0.365	0.385	0.371	0.394	0.355	0.386	0.362	0.392	0.358	0.390	0.381	0.399	0.404	0.419
192	0.102 (0.00)	0.242 (0.00)	0.110	0.258	0.405	0.411	0.403	0.412	0.401	0.417	0.398	0.418	0.396	0.414	0.435	0.431	0.451	0.449
336	0.146 (0.00)	0.303 (0.00)	0.150	0.313	0.437	0.433	0.420	0.422	0.420	0.429	0.430	0.427	0.391	0.420	0.480	0.452	0.471	0.465
720	0.228 (0.01)	0.381 (0.01)	0.232	0.392	0.448	0.462	0.422	0.444	0.425	0.447	0.442	0.457	0.430	0.457	0.499	0.488	0.565	0.538
Avg	0.138	0.286	0.148	0.301	0.414	0.423	0.404	0.418	0.400	0.420	0.408	0.424	0.394	0.420	0.449	0.443	0.473	0.468
ETTh2																		
96	0.125(0.00)	0.283(0.01)	0.175	0.312	0.276	0.332	0.269	0.332	0.273	0.334	0.268	0.328	0.257	0.326	0.297	0.347	0.305	0.361
192	0.159(0.00)	0.319(0.01)	0.232	0.361	0.344	0.377	0.328	0.377	0.334	0.371	0.329	0.375	0.309	0.361	0.373	0.394	0.389	0.411
336	0.196(0.01)	0.353(0.01)	0.259	0.380	0.377	0.406	0.353	0.396	0.353	0.398	0.368	0.409	0.302	0.366	0.410	0.426	0.420	0.434
720	0.241(0.01)	0.391(0.01)	0.278	0.411	0.406	0.435	0.383	0.425	0.371	0.412	0.381	0.417	0.372	0.416	0.411	0.433	0.436	0.454
Avg	0.180	0.336	0.236	0.366	0.351	0.387	0.333	0.382	0.333	0.379	0.336	0.382	0.310	0.367	0.372	0.400	0.387	0.415
ETTm1																		
96	0.216	0.336	0.215	0.330	0.283	0.330	0.285	0.343	0.280	0.331	0.272	0.334	0.286	0.338	0.325	0.361	0.305	0.358
192	0.309	0.416	0.294	0.404	0.320	0.354	0.324	0.366	0.321	0.357	0.310	0.358	0.327	0.366	0.375	0.389	0.343	0.380
336	0.462	0.513	0.418	0.484	0.357	0.377	0.353	0.385	0.347	0.378	0.352	0.384	0.362	0.389	0.405	0.412	0.380	0.402
720	0.548	0.564	0.525	0.548	0.418	0.412	0.408	0.419	0.395	0.411	0.383	0.411	0.414	0.421	0.466	0.447	0.441	0.438
Avg	0.384	0.457	0.363	0.442	0.345	0.368	0.343	0.378	0.336	0.369	0.329	0.372	0.347	0.379	0.393	0.402	0.367	0.395
ETTm2																		
96	0.362	0.414	0.270	0.376	0.162	0.244	0.165	0.254	0.164	0.254	0.161	0.253	0.162	0.244	0.180	0.261	0.179	0.268
192	0.446	0.480	0.421	0.472	0.223	0.285	0.220	0.292	0.221	0.291	0.219	0.293	0.220	0.291	0.246	0.306	0.242	0.312
336	0.526	0.557	0.552	0.584	0.274	0.320	0.268	0.326	0.273	0.326	0.271	0.329	0.272	0.326	0.319	0.352	0.292	0.344
720	0.623	0.641	0.590	0.616	0.354	0.373	0.350	0.380	0.355	0.378	0.352	0.379	0.358	0.379	0.405	0.401	0.380	0.397
Avg	0.575	0.541	0.479	0.523	0.253	0.305	0.251	0.313	0.253	0.312	0.251	0.313	0.253	0.312	0.287	0.330	0.273	0.330
ECL																		
96	0.327	0.409	0.142	0.298	0.129	0.217	0.128	0.223	0.129	0.220	0.131	0.224	0.127	0.217	0.143	0.233	0.132	0.228
192	0.472	0.473	0.159	0.318	0.147	0.234	0.146	0.240	0.147	0.238	0.152	0.241	0.144	0.234	0.158	0.248	0.155	0.249
336	0.581	0.572	0.171	0.333	0.164	0.253	0.163	0.258	0.157	0.249	0.160	0.248	0.162	0.256	0.178	0.269	0.171	0.266
720	0.694	0.669	0.194	0.352	0.186	0.275	0.200	0.292	0.198	0.285	0.192	0.298	0.199	0.289	0.218	0.305	0.206	0.299
Avg	0.519	0.531	0.167	0.325	0.157	0.245	0.159	0.253	0.158	0.248	0.159	0.253	0.157	0.245	0.174	0.264	0.166	0.261
Weather																		
96	0.069(0.00)	0.188(0.00)	0.077	0.195	0.148	0.194	0.147	0.196	0.141	0.182	0.147	0.201	0.150	0.199	0.166	0.208	0.163	0.212
192	0.135(0.00)	0.258(0.00)	0.144	0.269	0.195	0.239	0.191	0.238	0.188	0.228	0.189	0.234	0.194	0.240	0.217	0.253	0.207	0.253
336	0.221(0.01)	0.334(0.01)	0.230	0.341	0.252	0.282	0.241	0.277	0.240	0.273	0.262	0.279	0.246	0.281	0.282	0.300	0.256	0.291
720	0.392(0.02)	0.442(0.02)	0.413	0.445	0.323	0.331	0.313	0.329	0.312	0.322	0.304	0.315	0.319	0.333	0.356	0.351	0.330	0.340
Avg	0.205	0.306	0.216	0.313	0.229	0.262	0.223	0.260	0.220	0.251	0.225	0.257	0.227	0.263	0.255	0.278	0.239	0.274
Traffic																		
96	0.486	0.307	0.946	0.517	0.358	0.240	0.372	0.259	0.358	0.240	0.362	0.248	0.356	0.244	0.376	0.251	0.356	0.258
192	0.545	0.315	0.962	0.529	0.386	0.256	0.391	0.265	0.368	0.248	0.374	0.247	0.375	0.251	0.398	0.261	0.376	0.268
336	0.620	0.335	0.998	0.537	0.398	0.262	0.405	0.275	0.379	0.250	0.385	0.271	0.386	0.256	0.415	0.269	0.389	0.274
720	0.660	0.343	1.063	0.545	0.434	0.284	0.437	0.292	0.430	0.276	0.430	0.288	0.425	0.279	0.447	0.287	0.426	0.293
Avg	0.578	0.325	0.992	0.532	0.394	0.261	0.401	0.273	0.384	0.254	0.388	0.264	0.386	0.258	0.409	0.267	0.387	0.273
Exchange																		
96	0.102	0.226	0.102(0.00)	0.217(0.01)	0.088	0.210	- -		- -		- -		- -		- -		0.099	0.227
192	0.210	0.331	0.221(0.00)	0.326(0.01)	0.186	0.308	- -		- -		- -		- -		- -		0.202	0.326
336	0.396	0.471	0.348(0.02)	0.426(0.01)	0.350	0.428	- -		- -		- -		- -		- -		0.397	0.466
720	0.719	0.652	0.702(0.04)	0.641(0.02)	0.888	0.689	- -		- -		- -		- -		- -		0.947	0.740
Avg	0.387	0.430	0.343	0.404	0.378	0.409	- -		- -		- -		- -		- -		0.411	0.440
ILI																		
24	0.880	0.650	1.150(0.16)	0.771(0.09)	- -		- -		- -		1.285	0.727	- -		- -		- -	
36	1.355	0.815	1.123(0.09)	0.731(0.02)	- -		- -		- -		1.404	0.814	- -		- -		- -	
48	1.578	0.878	1.137(0.12)	0.751(0.06)	- -		- -		- -		1.281	0.794	- -		- -		- -	
60	2.143	1.061	1.395(0.09)	0.810(0.04)	- -		- -		- -		1.531	0.854	- -		- -		- -	
Avg	1.489	0.851	1.201	0.766	- -		- -		- -		1.375	0.797	- -		- -		- -	

tures both short-term fluctuations and periodic patterns in traffic data, achieving better performance across all baselines.

Figure 4 illustrates the 336-step forecasting performance of MSTN-Transformer, SOFTS, and iTransformer under the input-96-predict-336 setting. MSTN-Transformer achieves the lowest MSE and MAE and exhibits

Table 6: Long-term forecasting evaluation results on PEMS datasets (PEMS03, PEMS04, PEMS07, PEMS08). Prediction horizons $H \in \{12, 24, 48, 96\}$. Red/Blue: First/Second ranks. We evaluate each model five times and report the mean (standard deviation); results marked with underline are not significantly worse than the first rank (Wilcoxon signed-rank test with Holm–Bonferroni correction, $\alpha = 0.05$). iTransf: iTransformer.

Dataset/H	MSTN-Trans. (Ours)		MSTN-BiL. (Ours)		DARTS-TS (2025)		iTrans. (2024)		ModernTCN (2024)		PatchTST (2023)		TSMixer (2023)		TimesNet (2023)	
H-Metric	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
PEMS03																
12	0.010(0.002)	0.080(0.004)	0.018	0.102	0.066	0.171	0.069	0.175	0.112	0.221	0.079	0.187	0.075	0.187	0.085	0.192
24	0.019(0.002)	0.103(0.005)	0.032	0.132	0.097	0.206	0.098	0.209	0.173	0.281	0.124	0.235	0.113	0.238	0.110	0.216
48	0.032(0.004)	0.127(0.004)	0.060	0.178	0.152	0.257	0.448	0.416	0.307	0.395	0.223	0.319	0.195	0.320	0.168	0.263
96	0.048(0.004)	0.145(0.005)	0.072	0.185	0.234	0.331	1.215	0.831	1.041	0.779	0.368	0.425	0.266	0.380	0.242	0.321
Avg	0.027	0.114	0.046	0.149	0.137	0.241	0.458	0.408	0.408	0.419	0.199	0.291	0.162	0.281	0.151	0.248
PEMS04																
12	0.009(0.001)	0.071(0.003)	0.010	0.074	0.073	0.176	0.081	0.188	0.132	0.245	0.101	0.209	0.085	0.195	0.088	0.197
24	0.011(0.001)	0.080(0.003)	0.018	0.096	0.091	0.198	0.124	0.232	0.244	0.338	0.161	0.267	0.112	0.228	0.104	0.216
48	0.019(0.002)	0.096(0.008)	0.033	0.131	0.120	0.232	0.135	0.248	0.452	0.482	0.294	0.369	0.159	0.278	0.138	0.252
96	0.022(0.002)	0.099(0.004)	0.052	0.172	0.165	0.278	0.169	0.280	1.127	0.818	0.507	0.505	0.190	0.313	0.179	0.291
Avg	0.015	0.086	0.028	0.118	0.112	0.221	0.127	0.237	0.488	0.471	0.266	0.338	0.136	0.254	0.127	0.239
PEMS07																
12	0.007(0.001)	0.067(0.004)	0.010	0.074	0.060	0.155	0.066	0.164	0.085	0.196	0.076	0.180	0.070	0.177	0.083	0.183
24	0.008(0.001)	0.070(0.002)	0.015	0.091	0.081	0.180	0.087	0.190	0.127	0.245	0.127	0.234	0.105	0.221	0.101	0.204
48	0.013(0.001)	0.085(0.005)	0.020	0.106	0.113	0.218	0.892	0.764	0.267	0.380	0.238	0.325	0.157	0.265	0.133	0.236
96	0.013(0.001)	0.085(0.003)	0.031	0.123	0.156	0.262	0.972	0.789	0.736	0.673	0.394	0.432	0.268	0.342	0.211	0.308
Avg	0.010	0.077	0.019	0.099	0.102	0.204	0.504	0.477	0.304	0.374	0.209	0.293	0.150	0.251	0.132	0.233
PEMS08																
12	0.008(0.001)	0.069(0.004)	0.013	0.085	0.074	0.175	0.089	0.193	0.125	0.239	0.091	0.195	0.095	0.203	0.110	0.208
24	0.011(0.001)	0.082(0.003)	0.024	0.115	0.107	0.213	0.138	0.243	0.238	0.336	0.144	0.247	0.150	0.257	0.139	0.234
48	0.016(0.002)	0.091(0.006)	0.044	0.154	0.178	0.277	0.237	0.277	0.528	0.534	0.254	0.332	0.256	0.344	0.194	0.277
96	0.019(0.003)	0.100(0.005)	0.057	0.179	0.318	0.360	0.346	0.363	1.150	0.808	0.435	0.441	0.399	0.415	0.322	0.349
Avg	0.013	0.086	0.035	0.133	0.169	0.256	0.202	0.269	0.510	0.479	0.231	0.304	0.225	0.305	0.191	0.267

the realistic forecasting behavior, showing smoother trajectories with fewer abrupt fluctuations. Importantly, MSTN-Transformer consistently demonstrates better phase alignment with the ground truth, with predicted peaks and troughs occurring at nearly identical temporal locations, accurately preserving the periodic structure. Such temporal fidelity is critical for practical forecasting scenarios, where correct timing of events is often more important than exact amplitude matching. Although MSTN-Transformer produces slightly reduced peak amplitudes, this conservative behavior reflects an error-minimizing bias of MSE/MAE-based optimization, favoring stable trend estimation over aggressive extrapolation. In particular, in the late horizon, MSTN-Transformer adapts to the upward distribution shift of the ground truth and tracks the rising structural trend, whereas SOFTS, and iTransformer remain largely stationary and exhibit more erratic variations. Overall, these results demonstrate the effective long-horizon forecasting capability, robustness, and non-stationary adaptation ability of MSTN-Transformer. For clarity and readability, we visualize only three representative and competitive baselines, while quantitative results for all compared models are reported in Table 5.

4.6 Short-Term Forecasting Results

Short-term forecasting focuses on predicting immediate future values over a short time horizon, typically ranging from minutes to a few hours. This task is critical for real-time applications such as traffic management (congestion control, route optimization, incident detection), driving behavior analysis (vehicle trajectory prediction for collision avoidance), and autonomous navigation (anticipating nearby vehicle movements). Unlike long-term forecasting, which captures broader trends and seasonal patterns, short-term forecasting requires models to capture rapid fluctuations, sudden changes, and fine-grained temporal dynamics.

We evaluate short-term forecast performance using established protocols from the prior literature SimpleTM (Chen et al., 2025), TimeMixer (Wang et al., 2024) and SCINet (Liu et al., 2022a). We evaluate MSTN on four PEMS datasets (PEMS03, PEMS04, PEMS07, PEMS08) with $H = 12$. Table 7 reports MAE, MAPE, and RMSE on the original scale. MSTN-Transformer achieves better performance across all datasets,

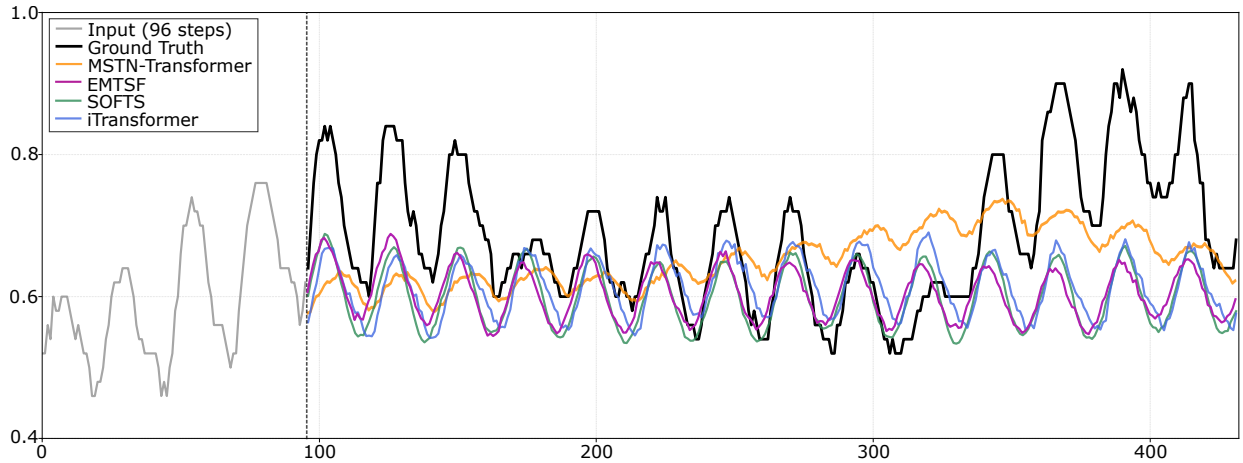

Figure 4: Visualization of weather predictions by different models under the input-96-predict-336 setting.

Table 7: Short-term forecasting results on PEMS datasets with $L = 96$ and $H = 12$. Metrics (MAE, MAPE, RMSE) are reported on the original scale after inverse transformation. Red/Blue: First/Second ranks. Results marked with underline are not significantly worse than the first rank according to the Wilcoxon signed-rank test with Holm–Bonferroni correction ($\alpha = 0.05$).

Dataset	MSTN-Tran. (Ours)	MSTN-BiL. (Ours)	SimpleTM (2025)	Time-Mixer (2024)	Cross-former (2023)	Patch-TST (2023)	Times-Net (2023)	DLinear (2023)	MICN (2023)	SCINet (2022)	FiLM (2022)	FED-former (2022)	Station-ary (2022)
PEMS03													
MAE	8.21(0.30)	8.65	14.86	14.63	15.64	18.95	16.41	19.70	15.71	15.97	21.36	19.00	17.64
MAPE	7.25(0.45)	7.94	14.79	14.54	15.74	17.29	15.17	18.35	15.67	15.89	18.35	18.57	17.56
RMSE	10.51(0.36)	11.22	23.58	23.28	25.56	30.15	26.72	32.35	24.55	25.20	35.07	30.05	28.37
PEMS04													
MAE	6.90(0.46)	7.00	18.74	19.21	20.38	24.86	21.63	24.62	21.62	20.35	26.74	26.51	22.34
MAPE	5.68(0.58)	5.82	12.11	12.53	12.84	16.65	13.15	16.12	13.53	12.84	16.46	16.76	14.85
RMSE	8.66(0.42)	8.84	30.46	30.92	32.41	40.46	34.90	39.51	34.39	32.31	42.86	41.81	35.47
PEMS07													
MAE	7.67(0.61)	7.86	20.25	20.57	22.54	27.87	25.12	28.65	22.28	22.79	28.76	27.92	26.02
MAPE	4.12(0.46)	4.38	8.55	8.62	9.38	12.69	10.60	12.15	9.57	9.41	11.21	12.29	11.75
RMSE	9.43(0.35)	9.57	33.06	33.59	35.49	42.56	40.71	45.02	35.40	35.61	45.85	42.29	42.34
PEMS08													
MAE	5.48(0.80)	5.57	14.92	15.22	17.56	20.35	19.01	20.26	17.76	17.38	22.11	20.56	19.29
MAPE	3.13(0.47)	3.25	9.36	9.67	10.92	13.15	11.83	12.09	10.76	10.80	12.81	12.41	12.21
RMSE	7.13(0.30)	7.21	23.80	24.26	27.21	31.04	30.65	32.38	27.26	27.34	35.13	32.97	38.62

followed by MSTN-BiLSTM. Compared to the recent SimpleTM (2025) baseline, MSTN-Transformer reduces MAE by 44.7% on PEMS03 (8.21 vs. 14.86), 63.2% on PEMS04 (6.90 vs. 18.74), 62.1% on PEMS07 (7.67 vs. 20.25), and 63.3% on PEMS08 (5.48 vs. 14.92). Compared to TimeMixer (2024), it achieves reductions of 43.9%, 64.1%, 62.7%, and 64.0% on the respective datasets. These results demonstrate that MSTN's multi-scale design—combining local pattern extraction with global dependency modeling is effective for short-term forecasting, effectively capturing both immediate fluctuations and periodic rush-hour patterns.

4.7 Time-Series Classification Results

Classification involves categorizing entire time series sequences. It is used to identify which type of event or pattern the sequence represents. For example, this could be the classification of a segment of sensor data as walking, running, falling, or collision activity, which is fundamental for automated monitoring and diagnosis systems. Performance is evaluated using the standard metric of classification accuracy.

Table 8: Comprehensive evaluation of MSTN across diverse time-series benchmarks. We evaluate each model five times and take the mean of the final results. **Red**/**Blue**: First/Second accuracy ranks. We evaluate each model five times and report the mean (standard deviation); results marked with underline are not significantly worse than the first rank (Wilcoxon signed-rank test with Holm–Bonferroni correction, $\alpha = 0.05$).

(a) Classification accuracy (%) comparison between MSTN (MSTN- Transformer and MSTN-BiLSTM) and SOTA baselines.

Dataset	MSTN BiL. (ours)	MSTN Trans. (ours)	GPT2 (6) (2023)	Times-(Net) (2023)	Light TS (2022)	DLinear (2023)	Flow. (2022)	ETS. (2022)	FED. (2022a)	Stat. (2022)	Auto. (2021)	Py. (2021a)	In. (2021)	Re. (2020)	Trans. (2017)
Ethanol	**41.51** (3.18)	**37.92 (3.71)**	31.9	35.7	29.7	32.6	33.8	28.1	31.2	32.7	31.6	30.8	31.6	31.9	32.7
FaceDetect	59.41 (1.26)	67.05 (1.94)	67.3	**68.6**	67.5	68.0	67.6	66.3	66.0	68.0	**68.4**	65.7	67.0	**68.6**	67.3
Handwriting	**91.20** (2.11)	**81.97** (4.34)	32.0	32.1	26.1	27.0	33.8	32.5	28.0	31.6	36.7	29.4	32.8	27.4	32.0
Heartbeat	**84.77** (2.78)	**82.42** (4.55)	76.1	78.0	75.1	75.1	77.6	71.2	73.7	73.7	74.6	75.6	80.5	77.1	76.1
JapaneseV	**99.31** (0.21)	98.63 (0.49)	98.6	98.4	96.2	96.2	98.9	95.9	98.4	**99.2**	96.2	98.4	98.9	97.8	98.7
PEMS-SF	**90.87** (1.78)	**89.92** (3.77)	82.1	89.6	88.4	75.1	83.8	86.0	80.9	87.3	82.7	83.2	81.5	82.7	82.1
SCP1	84.78 (3.37)	84.21 (3.47)	**93.2**	91.8	89.8	87.3	**92.5**	89.6	88.7	89.4	84.0	88.1	90.1	90.4	92.2
SCP2	54.27 (2.45)	53.58 (3.07)	**59.4**	**57.2**	51.1	50.5	56.1	55.0	54.4	**57.2**	50.6	53.3	53.3	56.7	53.9
SpokenArabic	**99.68** (0.22)	**100** (0.39)	99.2	99.0	**100**	81.4	98.8	**100**	**100**	**100**	**100**	99.6	**100**	97.0	98.4
UWave	**93.18**(1.43)	**92.37**(1.87)	88.1	85.3	80.3	82.1	86.6	85.0	85.3	87.5	85.9	83.4	85.6	85.6	85.6
Avg	**79.90**	**78.81**	72.79	73.67	70.43	65.53	72.85	70.96	70.66	72.73	71.07	70.75	72.13	71.52	71.90

(b) Generalizability Study: Performance comparison of MSTN-BiLSTM and MSTN-Transformer with TimesNet, PatchTST, and prior specialized methods (RF, GB, DT, kNN, SAE, etc.) across seven international cross-domain datasets. **Cyan**/**Violet**: First/Second inference time ranks.

				MSTN				TimesNet		PatchTST		Prior Work
Dataset	**Det.**	**Dom.**	**Cnt.**	**A.Block**	**Acc.(%)**	**Size**	**I.Time**	**Acc.(%)**	**I.Time**	**Acc.(%)**	**I.Time**	**Accuracy(%)**
Rodegast (Rodegast et al., 2024a)	Sim.	Hum.	DE.	BiLSTM	**99.25** (0.20)	1.34	**1.59**	99.18	5.07	98.64	4.80	91.00[RF,GB](Rodegast et al., 2024b)
				Transf.	**99.39** (0.14)	3.77	**4.03**					
Boubezoul(Boubezoul et al., 2020)	Real.	Hum.	FR.	BiLSTM	**99.16** (0.25)	1.03	**1.41**	93.00	4.20	91.37	4.68	91.59[DT](Elwy et al., 2023)
				Transf.	**97.67** (0.19)	3.54	**1.70**					
UCI-HAR(Reyes-Ortiz et al., 2013)	Act.	Hum.	IT.	BiLSTM	**96.41** (0.22)	1.68	**2.22**	91.38	45.30	93.21	4.36	83.35[Km,NB](Ismi et al., 2016)
				Transf.	**96.84** (0.12)	4.01	**3.63**					
PAMAP2(Reiss, 2012)	Phy.	Hum.	US.	BiLSTM	**99.73** (0.09)	1.64	**0.31**	95.13	**0.46**	98.02	3.21	90.00[kNN](Reiss & Stricker, 2012)
				Transf.	**99.89** (0.07)	4.13	1.90					
ActBeC.(Dissanayake et al., 2025)	Calf.	Anim.	IE.	BiLSTM	**93.00** (0.26)	3.95	**1.76**	90.65	8.53	62.45	2.20	84.00[RCCV](Dissanayake et al., 2025)
				Transf.	**92.85** (0.20)	4.14	**0.29**					
MetroPT3(Davari et al., 2021a)	Met.	Mech.	PT.	BiLSTM	**93.33** (0.16)	1.45	**0.07**	93.00	**0.09**	81.67	0.19	62.00[SAE](Davari et al., 2021b)
				Transf.	**95.00** (0.13)	3.98	0.69					
NASA(Saxena & Goebel, 2008)	Eng.	Mech.	US.	BiLSTM	**11.68**† (0.28)	1.12	**2.25**	15.56†	4.52	31.59†	2.31	–
				Transf.	**10.25**† (0.19)	3.62	**3.10**					

Abbreviations: Det: Details (Sim: Simulator PTW data, Real: Real PTW Data, Act: Activity human, Phy: Physical activity, Calf: Calf Behavior, Met: Metro predictive maintenance, Eng: Engine sensor), Dom: Domain (Hum: Human Safety, Anim: Animal welfare, Mech: Mechanical), Cnt: Country (IN: India, FR: France, DE: Germany, IT: Italy, US: USA, IE: Ireland, PT: Portugal), A.Block: Architecture block (Transf. Transformer), I.Time: Inference time, †RMSE values.

Table 8a presents the classification results across 10 UEA standard benchmark datasets. The proposed MSTN framework is compared against SOTA time-series classification methods, including TimesNet (Wu et al., 2023), GPT-2(6) (Zhou et al., 2023), FEDformer (Zhou et al., 2022), and LightTS (Zhang et al., 2022). MSTN demonstrates competitive performance across multiple datasets. MSTN-Transformer achieve perfect classification (100%) on SpokenArabic, with MSTN-BiLSTM achieves 99.68% accuracy. MSTN-BiLSTM achieves the better accuracy on multiple datasets: Ethanol (41.51%), Handwriting (91.20%), Heartbeat (84.77%), JapaneseVowels (99.31%), PEMS-SF (90.87%), and UWaveGestureLibrary (93.18%). MSTN-Transformer ranks second on multiple datasets. In terms of overall average accuracy, MSTN-BiLSTM ranks first (79.90%) and MSTN-Transformer ranks second (78.81%) among all baselines compared, confirming the consistent improvement of the proposed MSTN framework.

4.8 Generalizability Study

A robust model should not be a specialist in just one area; it should adapt to new tasks and data distributions across different data and domains. This experiment evaluates the ability of the MSTN to transfer its knowledge across diverse domains, from human activity recognition to mechanical fault prediction. As summarized in Table 8b, the proposed MSTN framework demonstrates generalization across seven international benchmark datasets spanning human safety, activity recognition, animal welfare, and mechanical

Table 9: Ablation on time series imputation (ETTh1) with different masking ratios (12.5%, 25%, 37.5%, 50%). **Red**/**Blue**: First/Second ranks.

MSTN Transformer Mask	Full MSE	Full MAE	w/o CNN MSE	w/o CNN MAE	w/o Transf. MSE	w/o Transf. MAE	w/o SE MSE	w/o SE MAE	w/o SDL MSE	w/o SDL MAE	w/o SGF MSE	w/o SGF MAE	w/o ETA MSE	w/o ETA MAE
12.5%	**0.002**	**0.013**	0.004	0.015	0.008	0.023	**0.003**	0.015	0.004	**0.014**	0.004	0.016	**0.003**	**0.014**
25%	**0.004**	**0.025**	0.007	0.030	0.015	0.047	0.006	**0.026**	0.006	0.028	**0.005**	**0.025**	**0.005**	0.027
37.5%	**0.007**	**0.037**	0.011	0.045	0.024	0.074	**0.009**	0.042	**0.009**	**0.041**	0.011	0.043	**0.006**	0.038
50%	**0.011**	**0.057**	0.015	0.060	0.035	0.101	0.015	0.062	**0.011**	0.060	**0.013**	0.062	**0.011**	**0.058**
MSTN BiLSTM	**Full**		**w/o CNN**		**w/o BiLSTM**		**w/o SE**		**w/o SDL**		**w/o SGF**		**w/o ETA**	
12.5%	**0.007**	**0.022**	0.009	0.025	0.010	0.026	**0.008**	0.024	0.009	**0.023**	0.009	0.024	**0.007**	0.024
25%	**0.012**	**0.042**	0.015	0.047	0.018	0.052	**0.013**	0.045	**0.012**	**0.043**	0.014	0.044	**0.013**	0.044
37.5%	**0.017**	**0.062**	0.021	0.069	0.026	0.075	0.019	0.066	0.019	0.064	**0.018**	**0.063**	**0.018**	0.064
50%	**0.022**	**0.081**	0.028	0.089	0.033	0.097	0.024	0.083	**0.023**	**0.085**	**0.023**	0.086	0.024	0.087

prognostics. Without domain-specific tuning, MSTN shows improved performance over state-of-the-art time series models—including TimesNet (Wu et al., 2023) and PatchTST (Nie et al., 2023)—as well as specialized prior methods (e.g., RF, GB, DT, kNN, SAE) reported in the Prior Work Accuracy column of Table 8b.

In human safety applications, MSTN-Transformer achieves 99.39% accuracy on German collision prediction data (Rodegast et al., 2024a), outperforming prior work (91.0%) and surpassing TimesNet (99.18%) and PatchTST (98.64%). For activity recognition, it achieves 96.84% on UCI-HAR (Reyes-Ortiz et al., 2013), surpassing TimesNet (91.38%) and PatchTST (93.21%). On fall detection, MSTN-Transformer achieves 99.39% and MSTN-BiLSTM 99.16% on Boubezoul, outperforming TimesNet (93.00%) and PatchTST (91.37%). On PAMAP2, MSTN-Transformer achieves 99.89%, exceeding TimesNet (95.13%) and PatchTST (98.02%). For animal welfare, MSTN-BiLSTM achieves 93.00% on calf behavior recognition (Dissanayake et al., 2025), outperforming TimesNet (90.65%) and PatchTST (62.45%). In mechanical prognostics, MSTN-Transformer achieves 95.00% on MetroPT3, surpassing TimesNet (93.00%) and PatchTST (81.67%). On the NASA Turbofan dataset (Saxena & Goebel, 2008), MSTN-Transformer achieves 10.25 RMSE, outperforming TimesNet (15.56) and PatchTST (31.59). MSTN-BiLSTM also shows improved performance.

Beyond accuracy, Table 8b highlights the structural impact of our ETA mechanism. While traditional models maintain full sequence complexity throughout deep layers, MSTN restricts this heavy computation to the initial encoding stage. Consequently, MSTN-BiLSTM achieves inference times of 1.59 ms on Rodegast (3.2× faster than TimesNet) and 0.07 ms on MetroPT3 (1.3× faster than TimesNet's 0.09 ms). On PAMAP2, MSTN-BiLSTM reaches 0.31 ms, significantly outperforming PatchTST (3.21 ms) by 10.4×. This confirms that MSTN effectively breaks the traditional trade-off between modeling capacity and latency.

4.9 Ablation Study

An ablation study systematically removes individual components from the full model to evaluate their contributions. Table 9 presents the ablation results for imputation task on ETTh1 dataset with varying mask ratios (12.5%, 25%, 37.5%, 50%). For MSTN-Transformer, six variants are tested: w/o CNN, w/o Transformer, w/o SE, w/o SGF, w/o SDL, and w/o ETA. For MSTN-BiLSTM, the variants include w/o CNN, w/o BiLSTM, w/o SE, w/o SGF, w/o SDL, and w/o ETA. The full MSTN-Transformer outperforms all variants, achieving the lowest MSE/MAE across all masks. The Transformer branch is most critical, with its removal causing the largest degradation (e.g., MSE increases from 0.002 to 0.008 at 12.5% mask, and from 0.011 to 0.035 at 50% mask), followed by the CNN branch whose removal also increases error. For MSTN-BiLSTM, the BiLSTM branch proves most important, with its removal causing substantial performance drops (e.g., MSE rises from 0.007 to 0.010 at 12.5% mask, and from 0.022 to 0.033 at 50% mask), while the CNN branch also demonstrates critical importance particularly at higher mask ratios. The SE and SDL modules show complementary benefits. Notably, removing ETA (w/o ETA) yields mixed results: for MSTN-Transformer, it degrades performance at low mask ratios (0.002 → 0.003 at 12.5%) but achieves best performance at 37.5% mask (0.006 MSE), outperforming the full model (0.007 MSE); for MSTN-BiLSTM, w/o ETA shows similar performance at 12.5% (0.007 vs 0.007) and causes degradation at higher masks (0.022 → 0.024 at 50%), confirming that ETA generally benefits both architectures.

Table 10: Ablation studies of MSTN-Transformer and MSTN-BiLSTM across different tasks. (a) Long-term Forecasting, (b) Forecasting on PEMS03 Dataset. **Red**/**Blue**: First/Second ranks.

(a) Ablation on Long-term forecasting horizons

MSTN Transformer	Full		w/o CNN		w/o Transf.		w/o SE		w/o SGF		w/o SDL		w/o ETA	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1														
96	**0.076**	**0.216**	0.117	0.275	0.142	0.308	0.087	0.234	0.099	0.251	0.158	0.331	**0.078**	**0.217**
192	**0.102**	**0.242**	0.173	0.346	0.123	0.278	0.118	0.265	0.121	0.267	**0.107**	**0.250**	0.110	0.258
336	**0.146**	**0.303**	0.168	0.338	0.201	0.376	0.151	0.318	**0.149**	**0.316**	0.228	0.408	0.149	0.317
720	**0.228**	**0.381**	0.242	0.423	0.268	0.446	**0.226**	0.404	0.257	0.438	0.314	0.492	**0.226**	**0.395**
ETTh2														
96	**0.125**	**0.283**	0.135	0.294	0.289	0.418	0.141	0.299	**0.130**	**0.289**	0.150	0.310	0.151	0.310
192	**0.159**	**0.319**	0.166	0.330	0.255	0.408	**0.161**	**0.321**	0.163	0.325	0.196	0.353	**0.161**	0.323
336	**0.196**	**0.353**	0.201	0.363	0.303	0.442	0.198	0.357	0.198	0.358	0.218	0.370	**0.197**	**0.355**
720	**0.241**	**0.391**	0.269	0.421	0.355	0.485	0.264	0.413	**0.262**	**0.412**	0.292	0.439	0.269	0.420
Exchange														
96	**0.102**	**0.226**	0.106	0.232	0.109	0.239	0.103	0.231	0.103	**0.229**	**0.101**	**0.229**	0.106	0.235
192	**0.210**	**0.331**	0.224	0.341	0.216	0.339	0.216	0.338	0.216	**0.336**	**0.215**	0.342	0.227	0.344
336	**0.396**	**0.471**	0.399	0.471	0.415	0.464	0.402	0.476	**0.396**	**0.472**	**0.398**	0.475	0.426	0.482
720	**0.719**	**0.652**	0.872	0.714	0.733	0.663	**0.717**	**0.652**	0.723	**0.655**	0.794	0.683	0.865	0.696
MSTN BiLSTM	**Full**		**w/o CNN**		**w/o BiLSTM**		**w/o SE**		**w/o SGF**		**w/o SDL**		**w/o ETA**	
ETTh1														
96	**0.101**	**0.240**	0.152	0.320	0.182	0.356	0.103	0.245	**0.101**	0.242	0.106	0.253	**0.099**	**0.238**
192	**0.110**	**0.258**	0.184	0.356	0.149	0.315	**0.112**	**0.260**	0.114	0.264	0.116	0.265	**0.112**	0.261
336	**0.150**	**0.313**	0.194	0.367	0.166	0.335	**0.153**	**0.320**	0.177	0.349	0.165	0.336	0.187	0.360
720	**0.232**	**0.392**	0.269	0.451	0.251	0.435	0.242	0.420	**0.234**	**0.403**	0.236	0.408	0.297	0.476
ETTh2														
96	**0.175**	**0.312**	0.256	0.407	0.278	0.422	0.259	0.403	0.276	0.409	**0.181**	**0.330**	0.299	0.421
192	**0.232**	**0.361**	0.263	0.413	0.295	0.433	0.276	0.415	**0.255**	**0.403**	0.280	0.421	0.321	0.447
336	**0.259**	**0.380**	0.290	0.434	0.316	0.451	**0.263**	**0.408**	0.291	0.431	0.287	0.429	0.285	0.426
720	**0.278**	**0.411**	0.368	0.492	0.356	0.486	**0.327**	**0.463**	0.351	0.483	0.332	0.468	0.352	0.483
Exchange														
96	**0.102**	**0.217**	0.108	0.237	0.106	0.236	0.104	0.229	0.105	0.232	0.104	0.228	**0.100**	**0.223**
192	**0.221**	**0.326**	0.238	0.354	0.247	0.368	0.226	0.337	0.226	0.339	**0.223**	0.333	0.224	**0.332**
336	**0.348**	**0.426**	0.415	0.484	0.390	0.471	**0.367**	0.451	0.371	0.457	**0.367**	**0.450**	0.370	0.461
720	**0.702**	**0.641**	0.835	0.731	0.725	0.655	0.717	0.657	0.726	0.665	0.720	0.661	**0.705**	**0.647**

(b) Ablation on PEMS03 Dataset.

MSTN Transformer	Full		w/o CNN		w/o Trans.		w/o SE		w/o SDL		w/o SGF		w/o ETA	
H	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
12	**0.010**	**0.080**	0.013	0.092	0.027	0.136	0.013	0.086	**0.010**	0.081	0.012	0.088	**0.009**	**0.079**
24	0.019	0.103	0.021	0.106	0.032	0.127	0.022	0.109	**0.014**	**0.083**	0.021	0.110	**0.017**	**0.089**
48	**0.032**	**0.127**	0.041	0.143	0.066	0.174	0.034	**0.127**	**0.030**	**0.129**	0.034	0.135	0.034	0.134
96	**0.048**	**0.145**	0.065	0.166	0.081	0.206	0.061	0.163	**0.049**	0.147	**0.048**	0.149	**0.049**	**0.146**
MSTN BiLSTM	**Full**		**w/o CNN**		**w/o BiL.**		**w/o SE**		**w/o SDL**		**w/o SGF**		**w/o ETA**	
12	**0.018**	**0.102**	0.020	0.106	0.027	0.136	0.019	0.103	0.020	0.103	**0.018**	0.104	**0.016**	**0.096**
24	**0.032**	0.132	0.038	0.140	0.040	0.143	0.036	0.136	**0.030**	**0.129**	0.034	0.133	**0.032**	**0.131**
48	**0.060**	0.178	0.072	0.181	0.075	0.186	0.061	0.178	**0.060**	**0.177**	0.062	0.180	**0.057**	**0.160**
96	**0.072**	**0.185**	0.078	0.192	0.081	0.206	**0.072**	**0.186**	0.073	0.189	0.074	0.187	0.073	0.187

Table 10 presents ablation studies on Long-term forecasting and forecasting on the PEMS03 dataset. Table 10a presents the Long-term forecasting ablation results for MSTN-Transformer and MSTN-BiLSTM on ETTh1, ETTh2, and Exchange datasets, across prediction horizons $H \in \{96, 192, 336, 720\}$.

For MSTN-Transformer, the full model achieves the best or second-best performance (ranked first/second in MSE/MAE) across nearly all prediction horizons and datasets. Removing the core Transformer module (w/o Transformer) leads to significant performance deterioration across all cases. For instance, on ETTh1 at horizon 96, MSE increases from 0.076 (full model) to 0.142 when the Transformer is ablated, highlighting its critical role in modeling complex temporal patterns. Similarly, the CNN branch (w/o CNN) is highly important; its removal causes noticeable degradation across all datasets, e.g., on ETTh1 MSE rising from 0.076 to 0.117 at H=96. The SE block (w/o SE) also contributes meaningfully, with the SE-ablated variant achieving

Table 11: Ablation studies of MSTN-Transformer and MSTN-BiLSTM across different tasks. (a) Short-term Forecasting on PEMS Dataset, (b) Classification on 10 UEA, (c) Generalizability Study. **Red**/**Blue**: First/Second ranks.

(a) Ablation study on PEMS03 short-term forecasting ($H = 12$). Metrics are MAE, MAPE, RMSE on original scale.

MSTN Transformer	**Full**	**w/o CNN**	**w/o Trans.**	**w/o SE**	**w/o SDL**	**w/o SGF**	**w/o ETA**
MAE	**7.71**	8.26	9.49	7.84	7.85	**7.75**	8.62
MAPE	6.97	7.85	8.09	**6.54**	**6.05**	6.99	8.93
RMSE	**9.90**	10.75	12.41	10.04	**8.70**	9.97	10.69
MSTN-BiLSTM	**Full**	**w/o CNN**	**w/o BiL.**	**w/o SE**	**w/o SDL**	**w/o SGF**	**w/o ETA**
MAE	**8.65**	9.12	9.49	9.63	9.09	9.16	**8.68**
MAPE	7.94	8.25	8.64	8.59	**7.75**	7.97	**7.47**
RMSE	**11.22**	11.76	12.41	12.34	12.36	12.09	**11.28**

(b) Ablation study on classification benchmarks.

MSTN Transformer	**Full**	**w/o CNN**	**w/o Transf.**	**w/o SE**	**w/o SDL**	**w/o SGF**	**w/o ETA**
PEMS-SF	**89.92**	86.25	85.11	89.26	85.44	**89.36**	81.82
JapaneseVowels	**98.63**	97.23	98.14	98.47	98.32	98.41	**98.44**
SpokenArabic	**100**	99.38	99.66	**99.84**	99.71	99.80	99.55
MSTN BiLSTM	**Full**	**w/o CNN**	**w/o BiLSTM**	**w/o SE**	**w/o SDL**	**w/o SGF**	**w/o ETA**
PEMS-SF	**90.87**	88.60	87.46	90.40	87.46	90.65	**90.91**
JapaneseVowels	**99.31**	96.97	99.05	99.14	**99.17**	99.00	98.44
SpokenArabic	**99.68**	99.00	99.28	**99.43**	99.37	99.33	99.20

(c) Ablation study on generalizability datasets.

MSTN Transformer	**Full**	**w/o CNN**	**w/o Transf.**	**w/o SE**	**w/o SDL**	**w/o SGF**	**w/o ETA**
PTW Simulation	**99.39**	98.89	98.73	99.04	99.08	98.93	**99.25**
PAMAP2	**99.89**	99.18	99.34	99.23	**99.65**	99.39	99.59
MSTN BiLSTM	**Full**	**w/o CNN**	**w/o BiLSTM**	**w/o SE**	**w/o SDL**	**w/o SGF**	**w/o ETA**
PTW Simulation	**99.25**	98.89	98.56	**99.11**	99.05	99.10	99.08
PAMAP2	**99.73**	99.39	99.35	99.62	**99.68**	99.56	99.51

competitive performance, ranking second on multiple horizons (e.g., ETTh1 at H=720 with MSE 0.226 vs. full model 0.228). The SDL (w/o SingleD) and SGF (w/o SGF) modules further enhance robustness, with their absence leading to consistent, though sometimes smaller, declines across datasets. Notably, removing ETA (w/o ETA) shows mixed effects: on ETTh1 at H=96, it achieves 0.078 MSE, closely matching the full model (0.076), while on Exchange at H=720, it causes degradation (0.865 vs. 0.719), indicating that ETA's benefit varies across datasets and architectures.

For MSTN-BiLSTM, the full model outperforms most ablated variants. Removing the BiLSTM layer (w/o BiLSTM) results in significant deterioration, particularly on ETTh1 where MSE at H=720 increases from 0.232 to 0.251. The CNN branch proves essential across all datasets, with ETTh1 MSE rising from 0.101 to 0.152 at H=96 when removed. The SE, SGF, and SDL modules provide complementary benefits, with their absence causing consistent performance drops. Interestingly, w/o ETA achieves the best performance on ETTh1 at H=96 (MSE 0.099 vs. full model 0.101), outperforming the full model, while showing degradation on other horizons (e.g., ETTh1 H=720: 0.297 vs. 0.232), suggesting that removing ETA benefits varies. This ablation study validates that each architectural component contributes meaningfully, with the full integrated model delivering robust performance across diverse forecasting tasks.

Table 10b presents ablation studies for MSTN-Transformer and MSTN-BiLSTM on PEMS data. Across both architectures, removing the SDL (w/o SDL) and witout ETA (w/o ETA), shows better performance in some horizons. w/o ETA achieves the better results at $H = 12$ for both Transformer (MSE=0.009, MAE=0.079) and BiLSTM (MSE=0.016, MAE=0.096). w/o SDL excels at $H = 24$ and $H = 48$ for Transformer, and at $H = 24$ for BiLSTM. The full models also show better performance while preserving a balance between

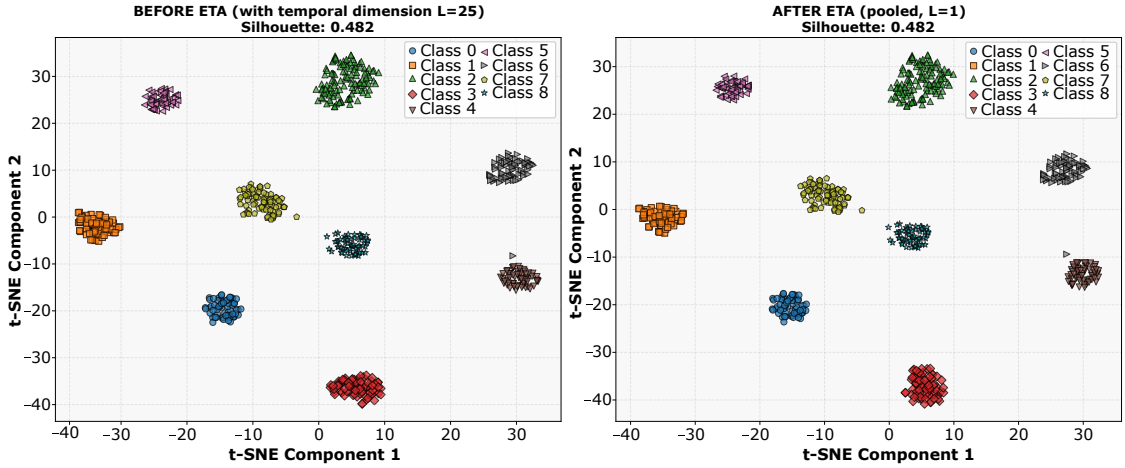

Figure 5: t-SNE visualization: Effect of ETA on JapaneseVowels dataset ($L = 25 \rightarrow 1$, 9 classes). Despite 25× temporal compression, the silhouette score remains unchanged (0.482 before vs. 0.482 after).

predictive accuracy and architectural complexity. These findings indicate that removing ETA and SDL are effective design choices that can be adapted based on specific dataset or tasks, whereas the full model with ETA offers a more balanced trade-off between performance and model complexity.

Table 11a presents the short term forecasting ablation results for MSTN-Transformer and MSTN-BiLSTM on the PEMS03 dataset. The Full MSTN-Transformer achieves the superior MAE of 7.71, though removing the Spatial Dependency Learning (SDL) module yields the lowest MAPE (6.05) and RMSE (8.70), suggesting the Transformer's attention mechanism efficiently handles spatial correlations independently. For the MSTN-BiLSTM, the Full model is the most robust, securing the top rank in MAE (8.65) and RMSE (11.22). Notably, removing the Sequential module (w/o Seq) causes the most significant performance drop (MAE=9.49) across both architectures. This underscores that while SDL or ETA impact specific metrics, the multi-scale temporal modeling provided by the Sequential module remains the indispensable core of the MSTN framework.

On classification benchmarks (PEMS-SF, JapaneseVowels, SpokenArabic) (Table 11b), the full MSTN-Transformer achieves 89.92%, 98.63%, and 100% accuracy respectively, while MSTN-BiLSTM achieves 90.87%, 99.31%, and 99.68% accuracy respectively. Ablation confirms that both branches are essential, with w/o ETA causing drops in most cases (e.g., MSTN-Transformer on PEMS-SF falls from 89.92% to 81.82%, MSTN-BiLSTM on JapaneseVowels drops from 99.31% to 98.44%), confirming the importance of ETA. Notably, w/o ETA surprisingly outperforms the full model for MSTN-BiLSTM on PEMS-SF (90.91% vs 90.87%) and achieves competitive results, indicating that ETA's benefit varies across datasets and architectures. The SE and SDL modules provide complementary gains.

On generalizability studies (PTW Simulation, PAMAP2) (Table 11c), both architectures show better performance, with MSTN-Transformer achieving 99.39%/99.89% and MSTN-BiLSTM achieving 99.25%/99.73% accuracy. The Transformer/BiLSTM branches are most critical, while w/o ETA achieves competitive results, (e.g., MSTN-Transformer on PTW Simulation drops from 99.39% to 99.25%; MSTN-BiLSTM on PAMAP2 drops from 99.73% to 99.51%), confirming ETA's importance for generalization. The SE and SDL modules provide consistent improvements across both architectures and datasets. In summary, the ablation study validates that each architectural component is essential. The Transformer/BiLSTM branch, convolutional branch, SE module, SDL, and ETA operate synergistically, with the full integrated model delivering the best results across diverse temporal datasets.

4.10 Comparing Representation Collapse: Models Maintaining L vs. ETA with t-SNE

To empirically verify that ETA preserves discriminative information despite collapsing the temporal dimension. We perform feature-space visualization using t-distributed Stochastic Neighbor Embedding (t-SNE)

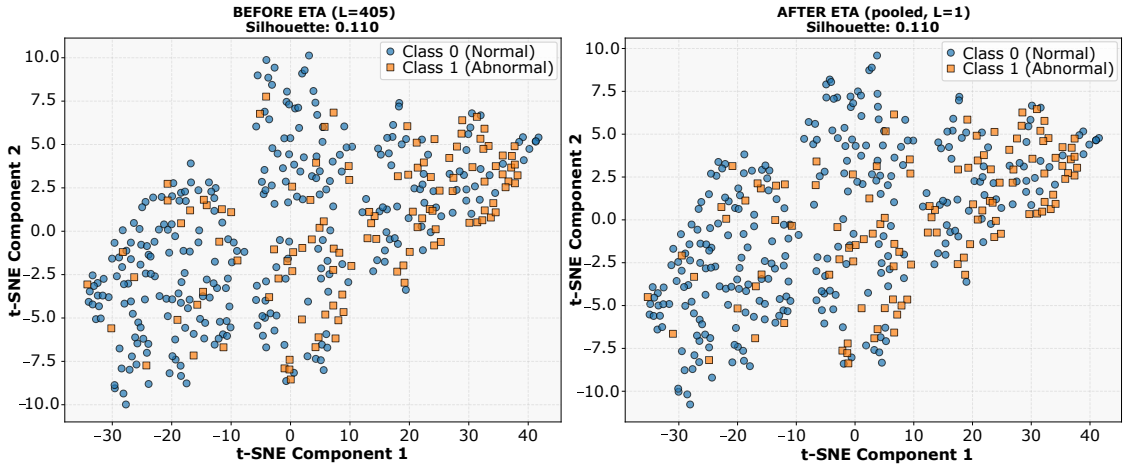

Figure 6: t-SNE visualization: Effect of ETA on Heartbeat dataset ($L = 405 \rightarrow 1$, binary). Despite 405× temporal compression, the silhouette score remains unchanged (0.110 before vs. 0.110 after).

visualization (Van der Maaten & Hinton, 2008) on features extracted before and after the ETA (pooling) operation on two diverse datasets: JapaneseVowels (multi-class, $L = 25$) and Heartbeat (binary, $L = 405$). We quantify cluster quality using the silhouette score (Rousseeuw, 1987).

Figure 5 shows the visualization of t-SNE on the JapaneseVowels dataset (9 classes, 640 samples, L=25). The silhouette score remains identical before and after ETA (0.482), demonstrating that 25× temporal compression preserves all class-discriminative information. Figure 6 shows the visualization of t-SNE on the Heartbeat dataset (binary classification, 409 samples, L=405). Remarkably, even at 405× compression, the silhouette score remains identical (0.110 before vs. 0.110 after), confirming that ETA preserves discriminative information even at extreme compression ratios.

These results conclusively demonstrate that ETA preserves discriminative information across diverse datasets, with identical silhouette scores before and after ETA (pooling). Models that maintain the full temporal dimension achieve the same feature quality as ETA's compressed representation ($L = 1$), indicating that there is no advantage in keeping L throughout the network. Furthermore, ETA achieves lossless temporal compression even at extreme ratios (405×), enabling $\mathcal{O}(1)$ complexity without sacrificing representation quality.

4.11 Model Complexity and Edge-AI Deployability

4.11.1 Complexity Analysis

In this analysis, we denote the length of the lookback window as L, the number of variate features (channels) as C, and the task-specific output dimension as K. We use C for channels to align with notation in common literature such as SOFTS (Han et al., 2024a), where K is often referred to as the forecasting horizon (H). We simplify the asymptotic complexity expressions by omitting the model dimension d, assuming $d \approx C$.

The MSTN framework addresses the fundamental trade-off in temporal modeling between representational capacity and computational efficiency. Although the proposed variants exhibit distinct encoding complexities, MSTN-BiLSTM maintains strict linearity with respect to sequence length and MSTN-Transformer incurring a quadratic cost. They both converge on a unified efficiency strategy: ETA. This is achieved because both the parallel CNN and sequence modeling branches utilize immediate temporal pooling to eliminate sequence-length dependence in subsequent layers. Specifically, both the parallel CNN and sequence modeling branches (Transformer/BiLSTM) independently condense their temporal features via Global Average Pooling and Sequence Mean Pooling, respectively (reducing the sequence dimension from L to 1). Following concatenation, the subsequent SGF, SE-Block, and SDL refinement stages operate on static fea-

ture vectors. This strategic design ensures that the refinement pipeline scales as $\mathcal{O}(C^2)$ with respect to channels but remains constant-time ($\mathcal{O}(1)$) with respect to the input length L, effectively isolating the heavy temporal computation to the initial stage.

Component-Wise Analysis of MSTN-Transformer: The complexity is dominated by the sequence encoding stage. First, the Multi-Scale CNN extracts local features with linear complexity $\mathcal{O}(CL)$. Second, the Transformer Encoder applies self-attention; as this requires computing an $L \times L$ attention matrix, it introduces the dominant quadratic term $\mathcal{O}(CL^2)$. Following this, Temporal Pooling aggregates the sequence with linear cost $\mathcal{O}(CL)$. Crucially, the Refinement stage operates on the pooled vector, scaling with channel interactions as $\mathcal{O}(C^2)$. Finally, the Projection Head maps features to the output with complexity $\mathcal{O}(CK)$. Summing these components, the total asymptotic complexity is $\mathcal{O}(CL^2+C^2+CK)$, which scales quadratically with sequence length L.

Component-Wise Analysis of MSTN-BiLSTM: In contrast, the MSTN-BiLSTM variant achieves strict linear efficiency with respect to sequence length. The BiLSTM Encoder processes the sequence recurrently without generating a global attention map, resulting in a linear complexity of $\mathcal{O}(CL)$. Similar to the Transformer variant, Temporal Pooling ($\mathcal{O}(CL)$) condenses the time dimension. The Refinement stage contributes $\mathcal{O}(C^2)$, and the Projection Head contributes $\mathcal{O}(CK)$. Since the encoder avoids quadratic operations in L, the total start-to-end complexity remains highly efficient at $\mathcal{O}(CL + C^2 + CK)$, which scales linearly with L.

This theoretical efficiency is empirically validated by the structural footprint of the model. As verified through our implementation, the MSTN-Transformer variant maintains a fixed core of $\sim$1.04 million parameters, while the MSTN-BiLSTM variant is highly optimized with only $\sim$0.40 million parameters. By ensuring a task-agnostic backbone, MSTN maintains consistent complexity across diverse datasets, meeting the requirement for architectural invariance. The transition from $\mathcal{O}(L)$ or $\mathcal{O}(L^2)$ in the encoding stage to $\mathcal{O}(1)$ in the refinement stage is the main reason for this low parameter count, allowing the model to perform effective feature recalibration without an additive parameter penalty.

4.11.2 Comparative Analysis and Edge-AI Deployability

Recent efficient architectures demonstrate varying complexity profiles: linear models such as DLinear (Zeng et al., 2022) achieve $\mathcal{O}(CL)$ complexity, while spectral approaches like FEDformer (Zhou et al., 2022) and TimesNet (Wu et al., 2023) achieve $\mathcal{O}(CL \log L)$ complexity. Following the notation in SOFTS (Han et al., 2024a), their model achieves $\mathcal{O}(CL + CH)$ complexity via stochastic pooling, while iTransformer ($\mathcal{O}(C^2 + CL + CH)$), PatchTST ($\mathcal{O}(CL^2 + CH)$), and Standard Transformers ($\mathcal{O}(N \cdot CL^2)$) incur heavier costs.

Against this landscape, the proposed MSTN-BiLSTM achieves strict linear temporal complexity $\mathcal{O}(CL+C^2+CK)$. Theoretically, this matches the optimal efficiency of SOFTS in the predominant real-world scenario where sequence length exceeds channel dimension ($L \gg C$). While SOFTS retains a theoretical advantage in high-dimensional, short-sequence settings ($C \gg L$) due to its purely linear channel dependence, such data structures are rare in long-term time series forecasting tasks. Meanwhile, MSTN-Transformer scaling as $\mathcal{O}(CL^2 + C^2 + CK)$ optimizes practical runtime by confining the quadratic term to a single encoder layer rather than a deep stack (N layers) as seen in iTransformer or standard Transformer architectures. This strategic design allows MSTN to balance the high-capacity modeling of Transformers with the latency requirements of Edge-AI, delivering SOTA performance with minimized computational overhead.

One thing to note is that despite the MSTN-BiLSTM's better theoretical scaling ($\mathcal{O}(L)$) than the MSTN-Transformer's $\mathcal{O}(L^2)$, the latter runs faster than the former (Table 12a). This phenomenon is due to the fundamental conflict between the model's structure and the way modern acceleration architectures process the data. The BiLSTM encoder is inherently recurrent, meaning that the calculation of the hidden state h_t is strictly dependent on the previous state h_{t-1}. This sequential data dependency prevents the hardware's parallel processing units from being fully utilized, forcing the computation into a slow, iterative loop (low hardware utilization). In contrast, the Transformer executes its $\mathcal{O}(L^2)$ self-attention via highly optimized (Vaswani et al., 2023), complete matrix parallelization. The core operation, the QK^T matrix multiplication, is performed simultaneously across thousands of processing elements. This high utilization of the hardware's parallel capacity, confined to a single encoder layer before pooling, ensures the Transformer's

Table 12: Comprehensive evaluation of MSTN across diverse time-series benchmarks. Inference time (IT, in ms) of MSTN compared to models including SOFTS, iTransformer, TimesNet and PatchTST. **Red**/**Blue**: First/Second ranks.

(a) Comparison of MSTN with the MSE and IT metrics reported in SOFTS (Han et al., 2024a) on the Traffic dataset ($L = 96$, $H = 720$).

	MSE	Perf. Gap	IT (ms)	Speed Gap
MSTN-Transformer	**0.020**	**1.00×**	**0.72**	**1.00×**
MSTN-BiLSTM	**0.087**	**4.30× ↓**	**3.20**	**4.44× ↓**
SOFTS	0.447	22.4× ↓	25.00	34.7× ↓
iTransformer	0.467	23.4× ↓	35.00	48.6× ↓
PatchTST	0.484	24.2× ↓	50.00	69.4× ↓
TS-Mixer	0.569	28.5× ↓	20.00	27.8× ↓
DLinear	0.645	32.3× ↓	10.00	13.9× ↓
Crossformer	0.589	29.5× ↓	150.00	208.3× ↓
Stationary	0.653	32.7× ↓	50.00	69.4× ↓
TimesNet	0.640	32.0× ↓	220.00	305.6× ↓
FEDformer	0.626	31.3× ↓	300.00	416.7× ↓

(b) Comprehensive evaluation of MSTN-Transformer across diverse time-series forecasting benchmarks (Wu et al., 2022).

Dataset	Horizon	Size (MB)	IT (ms)
Traffic	96,192,336,720	2.87	0.72
ETTh2	96,192,336,720	3.65	0.91
ETTm2	96,192,336,720	2.92	0.85
Weather	96,192,336,720	2.85	0.70
Exchange	96,192,336,720	2.96	0.80
ILI	24,36,48,60	2.51	0.63

(c) IT of MSTN variants compared to TimesNet and PatchTST across seven international datasets.

Dataset	MSTN	Size (MB)	IT (ms)	TimesNet IT	PatchTST IT
Rodegast (Rodegast et al., 2024a)	MSTN-Trans.	3.77	**4.03**	5.07	4.80
	MSTN-BiLSTM	1.34	**1.59**		
Boubezoul (Boubezoul et al., 2020)	MSTN-Trans.	3.54	**1.70**	4.20	4.68
	MSTN-BiLSTM	1.03	**1.41**		
UCI-HAR (Reyes-Ortiz et al., 2013)	MSTN-Trans.	4.01	**3.63**	45.30	4.36
	MSTN-BiLSTM	1.68	**2.22**		
PAMAP2 (Reiss, 2012)	MSTN-Trans.	4.13	1.90	**0.46**	3.21
	MSTN-BiLSTM	1.64	**0.31**		
ActBeCalf (Dissanayake et al., 2025)	MSTN-Trans.	4.14	**0.29**	8.53	2.20
	MSTN-BiLSTM	3.95	**1.76**		
MetroPT3 (Davari et al., 2021a)	MSTN-Trans.	3.98	0.69	**0.09**	0.19
	MSTN-BiLSTM	1.45	**0.07**		
NASA (Saxena & Goebel, 2008)	MSTN-Trans.	3.62	**3.10**	4.52	2.31
	MSTN-BiLSTM	1.12	**2.25**		

execution speed drastically outweighs the theoretical linearity of the BiLSTM's sequential cost, resulting in lower practical latency across high-throughput platforms.

Standard Transformers maintain the full sequence length L across a multi-layer stack of N blocks for self-attention operations, resulting in cumulative attention cost proportional to $N \cdot \mathcal{O}(CL^2)$. In contrast, MSTN applies attention in its encoder, then immediately pools ($L \to 1$). Standard Transformers maintain L across all layers, while MSTN eliminates L-dependence after encoding. Consequently, the remainder of the network operates on static feature vectors with $\mathcal{O}(1)$ complexity relative to L for operations involving sequence length. By eliminating the recursive application of attention across deep layers, MSTN reduces the sequence-dependent FLOPs. While MSTN-BiLSTM avoids quadratic attention costs entirely by employing BiLSTM encoding with $\mathcal{O}(CL)$ complexity. When combined with the remaining pipeline components (CNN, refinement, and projection), the total complexity is $\mathcal{O}(CL + C^2 + CK)$, which scales linearly with sequence length L.

Following the baseline results and inference time metrics reported in SOFTS (Han et al., 2024a) on the Traffic data set (lookback window $L = 96$, horizon $H = 720$), we compare our MSTN results with their

metrics reported in Table 12a. MSTN-Transformer demonstrates that architectural efficiency can surpass asymptotic predictions. MSTN-Transformer achieves 69× faster inference than PatchTST (0.72 ms vs. 50.00 ms). Furthermore, it outperforms models with better theoretical scaling, showing speedups of 34.7× over SOFTS and 305× over TimesNet. This practical gain, irrespective of the asymptotic class, is primarily delivered by the early temporal pooling of the MSTN architecture ($L \to 1$), a strategic design choice that dramatically reduces the overall FLOPs compared to deep, unpooled attention stacks. This performance represents a 22.4× accuracy improvement and 34.7× speedup over the SOFTS baseline, and a significant speed and accuracy advantage over all other models detailed in Table 12a. This improved efficiency enables substantially reduced memory usage and inference latency, positioning MSTN as a practical solution for deployment on constrained hardware platforms.

As summarized in Table 12b, MSTN demonstrates consistently efficient inference across six widely used long-term forecasting benchmarks. Despite the variation in dataset scale and temporal structure, MSTN maintains a compact model size in the range of 1.03–4.14 MB and sustains sub-millisecond latency across all horizons, including the longer 336 and 720-step settings. This stability in runtime behavior across heterogeneous datasets such as Traffic, ETTh2, ETTm2, Weather, Exchange, and ILI highlights the robustness of MSTN's dual-path, multi-scale feature extraction. The uniformly low inference cost emphasizes its suitability for real-time forecasting scenarios and deployment in resource-constrained environments where computational consistency is critical.

The inference analysis in Table 12c shows that MSTN maintains consistently low latency across seven heterogeneous datasets while maintaining a small footprint (1.03-4.14 MB). On the Rodegast dataset, MSTN-BiLSTM achieves 1.59 ms inference time, which is 3.2× faster than TimesNet (5.07 ms) and 3.0× faster than PatchTST (4.80 ms). On the Boubezoul dataset, MSTN-BiLSTM achieves 1.41 ms, outperforming TimesNet (4.20 ms) by 3.0× and PatchTST (4.68 ms) by 3.3×. On the ActBeCalf dataset, MSTN-Transformer achieves the fastest inference time of 0.29 ms, outperforming TimesNet (8.53 ms) by 29.4× and PatchTST (2.20 ms) by 7.6×. On the MetroPT3 dataset, MSTN-BiLSTM achieves an exceptionally low 0.07 ms, surpassing TimesNet (0.09 ms) and PatchTST (0.19 ms). These findings highlight the scalability and deployment efficiency of MSTN across both low-channel and high-dimensional sensor datasets, underscoring its suitability for latency-sensitive and edge-oriented time-series applications.

Finally, in terms of comparative complexity, MSTN shows an advantage over the current SOTA frameworks. While TIME-LLM (Jin et al., 2024) relies on a massive 6.6 billion parameter foundation model. The proposed MSTN-BiLSTM and MSTN-Transformer variants utilize fixed cores of only ∼0.40 million and ∼1.04 million parameters. This reduction in parameters compared to other models leads to better deployability of Edge-AI.

4.12 Limitations and Failure Cases

Although MSTN achieves state-of-the-art or competitive performance on most benchmarks, we analyze cases where specialized architectures outperform it, revealing important design trade-offs and current limitations. MSTN shows limitations on four forecasting datasets (ETTm1, ETTm2, ECL, Traffic) and three classification datasets (FaceDetect, SCP1, SCP2) where specialized architectures achieve better performance.

For Forecasting task, on ETTm1 and ETTm2, LLM4TS and DARTS-TS achieve better results (0.343–0.345 MSE on ETTm1; 0.251–0.253 MSE on ETTm2) compared to MSTN-Transformer (0.384 MSE on ETTm1; 0.299 MSE on ETTm2). These models leverage neural architecture search and large language model pre-training to capture the high-frequency patterns data more effectively than MSTN's general-purpose multi-scale design. On the Traffic dataset, HiMTM (0.384 MSE) outperforms all other latest baselines, including DARTS-TS, LLM4TS, MSTN-Transformer and MSTN-BiLSTM. This reflects HiMTM's specialized hierarchical design for complex multi-scale traffic data.

For the classification task, TimesNet's 2D modeling exceeds MSTN on FaceDetect with 68.6% accuracy (MSTN-BiLSTM: 59.4%, MSTN-Transformer: 67.1%), as it better captures the spatial-temporal relationships inherent in facial expression data by transforming 1D sequences into 2D representations. On SCP1 and SCP2, the pre-trained GPT2 model achieves 93.2% and 59.4% respectively, outperforming MSTN-BiLSTM and MSTN-Transformer. The language model benefits from its exposure to diverse sequential patterns during pre-training, enabling better transfer to self-regulation tasks that involve complex temporal dynamics.

These gaps reflect the fundamental trade-off in MSTN's design: prioritizing general-purpose multi-scale learning over specialized inductive biases. By focusing on broad applicability across heterogeneous scenarios, MSTN achieves better performance on most benchmarks but may be outperformed on datasets where explicitly optimized architectures have advantages. This trade-off reflects our design philosophy, positioning MSTN as a universal temporal model for diverse real-world applications rather than one tailored to particular datasets or domain-specific tasks. The modular architecture of MSTN allows practitioners to adapt it to their needs—for instance, by replacing the pooling-based ETA with a learnable temporal aggregation mechanism for high-frequency data, or by incorporating domain-specific preprocessing. These extensions are promising directions for future work.

4.13 Representational Limits of Linear Decoder

The linear decoder $\hat{Y}_{1:H} = W_f \mathbf{z}_{\text{final}} + b_f$ maps a latent vector of 192-dimensional to the horizon H. For $H = 720$, the compression ratio is $0.27 : 1$, which presents a theoretical bottleneck. However, three factors explain its empirical success in reconstructing complex dynamics including phase and high-frequency fluctuations:

(1) $\mathbf{z}_{\text{final}}$ is information-dense, capturing multi-scale features via CNN (local patterns), global dependencies via Transformer/BiLSTM, and adaptive fusion via SGF, SE, and SDL. This single vector encodes the essential characteristics of the entire input sequence.

(2) W_f learns 192 basis temporal patterns spanning different frequencies and phase shifts. The forecast is a linear combination $\hat{Y} = \sum_{i=1}^{192} \alpha_i \phi_i(t)$, where each basis $\phi_i(t)$ can represent different temporal dynamics—from low-frequency trends to high-frequency fluctuations—and coefficients α_i come from $\mathbf{z}_{\text{final}}$. This is analogous to learning a Fourier-like basis, but adapted to the data.

(3) Empirical Performance: Our model achieves 0.181 MSE on ETTh1 at $H = 720$ and 0.151 MSE on ETTm1 at $H = 720$ (ranked 1st among baselines, Table 5). Furthermore, ablation studies show that all components contribute at $H = 720$ (e.g., removing CNN increases MSE by 30%), demonstrating the richness of the latent representation.

For horizons $H > 1000$ or highly non-stationary series where the compression ratio exceeds the representational capacity of 192 bases, we recommend hierarchical decoding— a direction for future work.

5 Conclusions

This work introduces MSTN, a DL framework founded on a hierarchical multi-scale modeling principle and defined by its strategic use of ETA. MSTN integrates core components: a multi-scale convolutional encoder that constructs a hierarchical feature pyramid; a sequential modeling component for long-term global dependency modeling, empirically validated with BiLSTM and Transformer variants; and a SGF mechanism augmented with SE and SDL to enable dynamic, context-aware feature integration. Existing methods often rely on fixed-scale structural priors (e.g., patch-based tokenization, fixed frequency transformations, or frozen backbone architectures). These limitations have historically forced a compromise between scale diversity and inference speed. In contrast, our approach efficiently integrates convolutional local pattern extraction with long-term global dependency modeling, leveraging ETA via an adaptive SGF mechanism. This integrated design enables the model to capture multi-scale temporal features while maintaining computational efficiency, a combination we demonstrate yields SOTA performance.

Across extensive evaluations, MSTN demonstrates SOTA performance on 33 of 40 benchmark datasets spanning imputation, long term forecasting, short term forecasting, classification, and cross-dataset generalization. Notably, this improved performance is achieved with exceptional parameter efficiency. MSTN achieves this competitive performance with substantially fewer parameters than comparable SOTA methods: 0.40M (MSTN-BiLSTM) and 1.04M (MSTN-Transformer) parameters versus 6.6B for TIME-LLM (Jin et al., 2024) representing 16,500× reductions. MSTN-Transformer achieves a 22.4× reduction in prediction error (MSE) over the SOFTS model while being 34.7× faster in inference, all within a compact footprint, while MSTN-BiLSTM shows competitive performance. This dual advancement in both performance and efficiency is a distinguishing feature from prior work. This outcome, particularly the Transformer's better speed

despite its quadratic complexity, validates the MSTN's ETA strategy, showing that maximizing hardware parallelization in the encoding stage is often more critical for latency than maintaining theoretical linear complexity. The model also exhibits better cross-domain generalization, with its small model size (1.03–4.14 MB) making it suitable for edge deployment. Ablation studies confirm that each component is critical, with the Transformer/BiLSTM sequence modeling core providing a foundational capability and the multi-scale and fusion modules delivering significant competitive gains.

Although MSTN excels in standard forecasting regimes ($L \gg C$), we acknowledge that factorization-based models such as SOFTS retain a theoretical efficiency advantage in high-dimensional and short-sequence settings ($C \gg L$). However, since such datasets are rarely encountered in long-term time series forecasting tasks, this limitation has a negligible impact on practical deployability.

Future work will explore cross-variate attention mechanisms and large-scale pre-training to further enhance the capabilities of MSTN as a foundation model for time series understanding. Our work paves the way for efficient multi-scale temporal modeling under resource constraints, enabling real-time applications across diverse domains from healthcare to industrial monitoring.

Data and Code Availability

All datasets used in this study are publicly available from their respective sources as cited throughout the paper. The code supporting the results presented in this paper will be made publicly available upon publication.

Competing Interests

The authors declare no conflict of interest.

References

Musleh Alharthi, Kaleel Mahmood, Sarosh Patel, and Ausif Mahmood. Emtsf:extraordinary mixture of sota models for time series forecasting, 2025. URL https://arxiv.org/abs/2510.23396.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018, 2018. URL https://arxiv.org/abs/1811.00075.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018. URL http://arxiv.org/abs/1803.01271.

Abderrahmane Boubezoul, Fabien Dufour, Samir Bouaziz, and Stéphane Espié. Corrigendum to "dataset on powered two wheelers fall and critical events detection". *Data in Brief*, 30:105577, 2020. ISSN 2352-3409.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Trans. Intell. Syst. Technol.*, 16(3), April 2025. ISSN 2157-6904. doi: 10.1145/3719207. URL https://doi.org/10.1145/3719207.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oANkBaVci5.

Si-An Chen, Chun-Liang Li, Sercan O Arik, Nathanael Christian Yoder, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecast-ing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wbpxTuXgm0.

Yu Chen and PeMS data. Pems, 2025. URL https://test.ieee-dataport.org/documents/pems-3. California Traffic Network Datasets: PEMS03, PEMS04, PEMS07 and PEMS08 for Traffic Flow Forecasting.

Narjes Davari, Bruno Veloso, Rita Ribeiro, and Joao Gama. MetroPT-3 Dataset. UCI Machine Learning Repository, 2021a. DOI: https://doi.org/10.24432/C5VW3R.

Narjes Davari, Bruno Veloso, Rita P. Ribeiro, Pedro Mota Pereira, and João Gama. Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, 2021b. doi: 10.1109/DSAA53316.2021.9564181.

Difan Deng and Marius Lindauer. Optimizing time series forecasting architectures: A hierarchical neural architecture search approach. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Ym2wqojm4e.

Oshana Iddi Dissanayake, Sarah E. McPherson, Joseph Allyndrée, Emer Kennedy, Pádraig Cunningham, and Lucile Riaboff. Actbecalf: Accelerometer-based multivariate time-series dataset for calf behavior classification. *Data in Brief*, 60:111462, 2025. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2025.111462. URL https://www.sciencedirect.com/science/article/pii/S2352340925001945.

Luo donghao and wang xue. ModernTCN: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series, 2024. URL https://arxiv.org/abs/2401.03955.

Fatema Elwy, Raafat Aburukba, A. R. Al-Ali, Ahmad Al Nabulsi, Alaa Tarek, Ameen Ayub, and Mariam Elsayeh. Data-driven safe deliveries: The synergy of iot and machine learning in shared mobility. *Future Internet*, 15(10), 2023. ISSN 1999-5903.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4652–4663, 2019.

Ayuningtyas Hari Fristiana, Syukron Abu Ishaq Alfarozi, Adhistya Erna Permanasari, Mahardhika Pratama, and Sunu Wibirama. A survey on hyperparameters optimization of deep learning for time series classification. *IEEE Access*, 12:191162–191198, 2024. doi: 10.1109/ACCESS.2024.3516198.

Aleksandra Gruca, Federico Serva, Llorenç Lliso, Pilar Rípodas, Xavier Calbet, Pedro Herruzo, Jiří Pihrt, Rudolf Raevskyi, Petr Šimánek, Matej Choma, et al. Weather4cast at neurips 2022: Super-resolution rain movie prediction under spatio-temporal shifts. In *NeurIPS 2022 Competition Track*, pp. 292–313. PMLR, 2022.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion, 2024a. URL https://arxiv.org/abs/2404.14197.

Wenyong Han, Tao Zhu, Liming Chen, Huansheng Ning, Yang Luo, and Yaping Wan. Mcformer: Multivariate time series forecasting with mixed-channels transformer. *IEEE Internet of Things Journal*, 11(17):28320–28329, 2024b. doi: 10.1109/JIOT.2024.3401697.

Yangdong He and Jiabao Zhao. Temporal convolutional networks for anomaly detection in time series. *Journal of Physics: Conference Series*, 1213(4):042050, 2019. doi: 10.1088/1742-6596/1213/4/042050.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Dewi Pramudi Ismi, Shireen Panchoo, and Murinto Murinto. K-means clustering based filter feature selection on high dimensional data. *International Journal of Advances in Intelligent Informatics*, 2:38–45, 2016. URL https://api.semanticscholar.org/CorpusID:43897444.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models, 2024. URL https://arxiv.org/abs/2310.01728.

Akhil Kadiyala and Ashok Kumar. Multivariate time series models for prediction of air quality inside a public transportation bus using available software. *Environmental Progress & Sustainable Energy*, 33(2): 337–341, 2014.

E. G. Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas, and A. G. Bakirtzis. Application of time series and artificial neural network models in short-term forecasting of pv power generation. In *2013 48th International Universities' Power Engineering Conference (UPEC)*, pp. 1–6, 2013. doi: 10.1109/UPEC.2013.6714975.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Olaf Köllé. Wetterstation. weather. Technical report and dataset, Max-Planck-Institut für Biogeochemie (BGC Jena), Germany, 2025. URL https://www.bgc-jena.mpg.de/wetter/.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks, 2018. URL https://arxiv.org/abs/1703.07015.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021. doi: 10.1098/rsta.2020.0209.

Wenxie Lin. California traffic network datasets, 2025. PEMS03, PEMS04, PEMS07, PEMS08 traffic flow datasets from Caltrans PeMS.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction, 2022a. URL https://arxiv.org/abs/2106.09305.

Yijing Liu, Qinxian Liu, Jian-Wei Zhang, Haozhe Feng, Zhongwei Wang, Zihan Zhou, and Wei Chen. Multivariate time-series forecasting with temporal polynomial graph neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=pMumil2EJh.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting, 2024. URL https://arxiv.org/abs/2310.06625.

Kiran Madhusudhanan, Shayan Jawed, and Lars Schmidt-Thieme. Hyperparameter tuning mlp's for probabilistic time series forecasting. Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-981-97-2265-5.

Mohammad Amin Morid, Olivia R. Liu Sheng, and Joseph Dunbar. Time series prediction using deep learning methods in healthcare. 14(1), January 2023. ISSN 2158-656X. doi: 10.1145/3531326. URL https://doi.org/10.1145/3531326.

Tam Nguyen, Tan M. Nguyen, and Richard G. Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers, 2023. URL https://arxiv.org/abs/2211.14730.

Satoshi Noguchi and Yoshinobu Kawahara. Wavy transformer. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, NeurIPS '25, 2025.

Evandro S. Ortigossa and Eran Segal. Seg-moe: Multi-resolution segment-wise mixture-of-experts for time series forecasting transformers, 2026. URL https://arxiv.org/abs/2601.21641.

Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective, 2025. URL https://arxiv.org/abs/2502.10721.

Attila Reiss. PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pp. 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C54S4K.

M. Rodegast et al. Motorcycle collision dataset, 2024a. URL https://darus.uni-stuttgart.de/dataset.xhtml?persistentId=doi:10.18419/darus-3301.

Philipp Rodegast, Steffen Maier, Jonas Kneifl, and Jörg Fehr. On using machine learning algorithms for motorcycle collision detection. *Discover Applied Sciences*, 6(6):326, 2024b. ISSN 3004-9261.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7.

Abhinav Saxena and Kai Goebel. Nasa turbofan engine degradation simulation data set, 2008. URL https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/. NASA Ames Prognostics Center of Excellence.

Chao Shang, Jie Chen, and Jinbo Bi. Discrete graph structure learning for forecasting multiple time series, 2021. URL https://arxiv.org/abs/2101.06861.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting, 2022. URL https://arxiv.org/abs/2202.01381.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, 2022. URL https://arxiv.org/abs/2106.13008.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis, 2023. URL https://arxiv.org/abs/2210.02186.

Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective, 2023. URL https://arxiv.org/abs/2311.06190.

Wenzhen Yue, Ruohao Guo, Ji Shi, Zihan Hao, Shiyu Hu, and Xianghua Ying. vlinear: A powerful linear model for multivariate time series forecasting, 2026. URL https://arxiv.org/abs/2601.13768.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022. URL https://arxiv.org/abs/2205.13504.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures, 2022. URL https://arxiv.org/abs/2207.01186.

Yitian Zhang, Liheng Ma, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. Multi-resolution time-series transformer for long-term forecasting, 2024. URL https://arxiv.org/abs/2311.04147.

Shubao Zhao, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Qingsong Wen, and Yi Wang. Himtm: Hierarchical multi-scale masked time series modeling with self-distillation for long-term forecasting, 2024. URL https://arxiv.org/abs/2401.05012.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting, 2022. URL https://arxiv.org/abs/2201.12740.

Tian Zhou, PeiSong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all:power general time series analysis by pretrained lm, 2023. URL https://arxiv.org/abs/2302.11939.