Mamba-LSTM-Attention (MLA): A Hybrid Architecture for Long-Term Time Series Forecasting with Cross-Scale Stability

Fusheng Chen , Chong Fo Lei , Te Guo , Chia Wei Chu *

Posted Date: 22 January 2026

doi: 10.20944/preprints202601.1709.v1

Keywords: multivariate time series forecasting; Mamba; state space models; long-term temporal dependencies; hybrid neural architectures; smart grid load forecasting

Article

Mamba-LSTM-Attention (MLA): A Hybrid Architecture for Long-Term Time Series Forecasting with Cross-Scale Stability

Fusheng Chen [1], Chong Fo Lei [1,2], Te Guo [1] and Chiawei Chu [1,*]

[1] Faculty of Data Science, City University of Macau, Avenida Padre Tomás Pereira Taipa, Macao, China; ORCiD: 0000-0002-9839-1024, ORCiD: 0009-0002-5626-1501, ORCiD: 0000-0002-1334-9848

[2] Faculty of Creative Tourism and Intelligent Technologies, Macao University of Tourism, Colina de Mong-Ha, Macao, China; Email: fordlei@utm.edu.mo; ORCiD: 0000-0002-6756-0470

* Correspondence: cwchu@cityu.edu.mo

Abstract

Forecasting long-term time series (LTSF) requires a delicate balance between capturing global dependencies and preserving granular local dynamics. Although State Space Models (SSMs) and Transformers have been successful, a technical challenge remains, in which individual paradigms often fail to perform well over extended horizons due to the difficulty of simultaneously achieving linear-time efficiency and high-fidelity local refinement at the same time. This study introduces Mamba-LSTM-Attention (MLA), a novel hybrid architecture featuring a cascaded topology designed to bridge these dimensions. In this work, the core innovation is the hierarchical feature evolution mechanism, in which a Mamba module serves first as an efficient global encoder to capture long-range periodic trends at linear complexity. Following this, gated LSTM units are used to refine the algorithm at the micro-scale in order to filter noise and characterize non-linear local fluctuations. Lastly, a multi-head attention mechanism performs a dynamic feature re-weighting in order to focus on key historical signals. Systematic evaluations across four multivariate benchmark datasets demonstrate that MLA achieves exceptional cross-step forecasting stability. Most notably, on the ETTh1 dataset, MLA maintains a remarkably narrow Mean Squared Error (MSE) fluctuation range (0.127) as the forecasting horizon extends from T=96 to T=720. This empirical evidence confirms that the integrated Mamba module effectively mitigates the error accumulation typically encountered by vanilla LSTMs. While the current implementation faces an information bottleneck due to a single-point projection decoding strategy, the ablation studies (revealing a 19.76% MSE surge upon LSTM removal) validate the combination of the proposed architecture. This work establishes a robust framework for hybrid SSM-RNN modeling and a clear path for future performance enhancements through sequence-to-sequence mechanisms.

Keywords: multivariate time series forecasting; Mamba; State Space Models; long-term temporal dependencies; hybrid neural architectures; smart grid load forecasting

1. Introduction

Time series forecasting represents a core component of data science, providing critical decision support across multiple industries, such as power load management, traffic flow control, and meteorological warning systems [1]. Driven by advancements in data acquisition and storage, research focus has shifted from short-term data toward Long-Term Time Series Forecasting (LTSF), which is essential for identifying deep periodicities and informing long-range strategic planning. However, modern LTSF tasks are increasingly characterized by multivariate sequences that exhibit high dimensionality and complex spatiotemporal correlations. For instance, in a smart grid scenario, a model must not only predict load demands across extended horizons but also simultaneously

 doi:10.20944/preprints202601.1709.v1

process multi-source variables such as weather indices and historical voltage fluctuations, which are often subject to non-linear coupling [28,32]. This dual expansion in both temporal and feature dimensions imposes more rigorous demands on both the representational capacity and computational efficiency of predictive models.

Despite the importance of these multivariate forecasting tasks [28,31], existing architectures face a difficult trade-off between global awareness and local precision. This challenge arises because expanding the receptive field to capture long-range dependencies as seen in Transformers often leads to the "smoothing out" of granular, high-frequency signals. Conversely, while classic deep learning models like LSTM excel at characterizing immediate non-linear dynamics, they remain susceptible to "receptive field decay" and signal attenuation as the forecasting horizon T extends, frequently resulting in "performance collapse" [2]. Even though the self-attention mechanism in Transformers achieves global context breakthroughs, its quadratic computational complexity presents severe memory bottlenecks for ultra-long sequences [3–6]. Consequently, a structural gap remains for an architecture that can simultaneously achieve linear-time efficiency and high-fidelity local refinement.

To overcome these constraints, State Space Models (SSMs), particularly the Mamba architecture, have garnered widespread attention by achieving linear computational complexity through a selective state space mechanism [7,8]. Mamba enables the adaptive filtering of redundant information while retaining critical historical states, often matching or exceeding Transformer performance in long-sequence tasks [9,10]. However, while Mamba has seen significant success in language and genomics [11], its application in LTSF remains exploratory, and a sole reliance on SSMs may overlook the fine-grained local patterns essential for high-precision forecasting [12,13]. Consequently, diverse architectures within a forecasting model framework may offer distinct and complementary strengths. For instance, Mamba is capable of modeling expansive long-range dependencies [15], LSTM exhibits unique robustness in characterizing local temporal dynamics [16], and Attention mechanisms excel at capturing significant correlations within specific windows [14]. Although recent studies suggest that hybridizing Mamba with Transformers can enhance representational capacity [17,18], existing work has not yet fully explored how to organically integrate these three components to synergistically optimize multivariate long-term forecasting. There is an urgent need for a unified paradigm that bridges the gap between global context awareness and local temporal refinement.

In order to address these challenges, this paper proposes a novel hybrid deep learning architecture Mamba-LSTM-Attention (MLA), whichspecifically engineered to harmonize the distinct advantages of disparate paradigms into a unified framework. The structural core of the MLA algorithm resides in its triple-complementary, hierarchical design, which facilitates a progressive feature evolution process. This hybrid model starts with a Mamba module, which is deployed as the foundational feature extractor, leveraging its selective state space mechanism to efficiently encode expansive global dependencies at linear complexity. Subsequently, gated LSTM units are employed to perform fine-grained refinement of local temporal patterns, ensuring the precise characterization of short-term non-linear fluctuations. Finally, a multi-head attention mechanism is integrated to dynamically re-weight these refined features, bolstering the model's perception of significant historical signals and pivotal time steps. MLA achieves superior structural stability and robust predictive performance on multivariate time series data by strategically balancing between global context awareness and local temporal nuances.

2. Literature Review

2.1. The Current Development of Time Series Forecasting

The field of time series forecasting has undergone a significant evolution, transitioning from traditional statistical methods to sophisticated modern deep learning models. Early research primarily relied on linear models such as Autoregressive Integrated Moving Average (ARIMA) [19] and Exponential Smoothing. While computationally efficient and highly interpretable, these methods struggle to capture the non-linear dynamics and complex long-range dependencies inherently

 doi:10.20944/preprints202601.1709.v1

present in real-world time series data [20]. With the growth of computational power and the advent of the big data era, deep learning models have gradually become the mainstream for time series analysis. Among these, Recurrent Neural Networks (RNNs) and their advanced variants—Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)—effectively model short-term dependencies through internal hidden state transition mechanisms [21,22]. LSTM, in particular, leverages its unique gating mechanisms (forget, input, and output gates) to mitigate the vanishing gradient problem inherent in vanilla RNNs. It has demonstrated exceptional performance across diverse time series forecasting tasks, including speech recognition, natural language processing, and applications in finance, meteorology, and energy [7,23]. Numerous empirical studies have shown that LSTM typically outperforms traditional methods like ARIMA when processing sequence data characterized by complex fluctuations and non-linear patterns [1,24].

Despite these successes, LSTM models still face challenges in handling extremely long sequences, as their capacity to capture long-range dependencies remains limited. Consequently, the Attention Mechanism was introduced to temporal modeling. This mechanism allows a model to dynamically focus on different segments of the input sequence and assign varying weights, thereby better capturing information from critical time steps [2]. Researchers have proposed various hybrid architectures combining LSTM and Attention (e.g., Attention-LSTM), which further enhance predictive performance by bolstering the model's perception of significant historical moments [3–5]. The emergence of the Transformer architecture marked another milestone in time series forecasting. Its core Self-Attention mechanism enables the parallel computation of relationships across all time steps, vastly improving the model’s ability to capture global context and long-range dependencies [6]. Transformer-based models, such as Informer and Autoformer, have achieved state-of-the-art (SOTA) performance on several long-term time series forecasting benchmarks [7]. However, Transformers are notorious for their quadratic computational complexity relative to sequence length, leading to substantial computational and memory overhead when processing ultra-long sequences [8].

To address the computational bottlenecks of Transformers, State Space Models (SSMs) have recently regained significant attention. Structured State Space Sequence Models (S4) and the subsequent Mamba model, through selective state space mechanisms, have demonstrated performance comparable or superior to Transformers in long-sequence tasks such as language, audio, and genomics, all while maintaining linear computational complexity [9]. The Mamba model adaptively filters redundant information and preserves key historical states, offering a novel solution for efficient long-sequence modeling [10]. Recently, hybrid models have emerged as a research hotspot, as investigators seek to integrate the strengths of diverse architectures to compensate for the limitations of individual models. Examples include combining LSTM with Attention to simultaneously capture local patterns and critical time points [11,12], integrating Transformers with LSTM to balance global awareness and local refinement [13], and exploring the integration of the emerging Mamba architecture with other components to achieve a synergy of efficiency and accuracy in long-sequence forecasting [14]. These studies provided valuable insights into constructing more robust predictive models.

2.2. Multivariate Time Series Data in Long-Term Time Series Forecasting (LTSF)

The complexity of Multivariate Long-term Time Series Forecasting (LTSF) arises from the coupled dynamics of temporal evolution and inter-series correlations [4]. Unlike univariate scenarios, multivariate datasets (e.g., electricity grids or traffic networks) require models to capture "spatial-temporal duality," where the state of one variable is often conditioned on the historical trajectories of others. Early deep learning attempts, such as LSTNet [3] and DeepAR [5], utilized hybrid CNN-RNN or autoregressive structures to decode these patterns, but their efficacy diminishes as the forecasting horizon T extends, often leading to cumulative error propagation. The current research landscape is dominated by a debate over how to represent multivariate channels. A challenge was posed by Zeng et al. (2023), who introduced DLinear [25], arguing that complex Transformer-based spatial-temporal

 doi:10.20944/preprints202601.1709.v1

embeddings might inadvertently destroy temporal order, and that simple linear mappings can often yield superior results by maintaining channel-specific trends. In response, Nie et al. (2023) proposed the Channel-Independence (CI) strategy in PatchTST [26], treating each variable as an isolated sequence to enhance robust feature extraction. While CI mitigates noise interference between heterogeneous variables, it inherently fails to model the synergistic interactions that define complex multivariate systems.

To address the limitations of channel independence, several "Channel-Mixing" architectures have been developed. Zhang and Yan (2023) proposed Crossformer [27], which utilizes a two-stage attention mechanism to explicitly capture dependencies across both time and dimension axes. Furthermore, Liu et al. (2024) introduced iTransformer [29], which inverts the traditional Transformer structure by treating each variable as a token, thereby allowing self-attention to learn global multivariate correlations. Other significant approaches include MTGNN [29], which utilizes graph neural networks to learn hidden structural associations, and TSMixer [30], an all-MLP architecture designed for efficient mixing of temporal and feature information. Wu et al. (2023) also proposed TimesNet [31], which transforms 1D time series into 2D variations to better capture multi-periodicity in multivariate signals.

However, the simultaneous pursuit of high-dimensional representational fidelity and computational scalability remains an open problem. While MLP-based models like TiDE [33] and convolutional structures like SCINet [33] offer efficiency, they often lack the adaptive long-range context awareness. Recent explorations into State Space Models, such as Time-Mamba [34], suggest that Mamba-based frameworks can handle multivariate data with linear complexity. Nevertheless, as highlighted in the present study, these models still struggle to balance expansive global encoding with the granular, local temporal refinement provided by LSTM [20] and the dynamic weighting of Attention [22] mechanisms.

While LSTM, Attention mechanisms, Transformers, and State Space Models each play a vital role in time series forecasting with complementary advantages, the integration of Mamba's long-range modeling efficiency, Attention's dynamic focusing capability, and LSTM's local temporal characterization remains a relatively unexplored area. Due to this reason, this study aims to contribute to long-term time series forecasting research by proposing a MLA framework that integrates Mamba, LSTM, and Attention hierarchically.

3. Methodology

This study proposes a novel deep learning architecture, Mamba-LSTM-Attention (MLA), aimed at addressing pivotal challenges in Long-term Time Series Forecasting (LTSF), including capturing long-range dependencies, modeling local temporal dynamics, and performing effective global feature re-weighting. By cascading three powerful deep learning components, the Mamba state space model, LSTM, and the attention mechanism, the MLA model forms a network structure capable of simultaneously processing both long-term and short-term dependencies.

 doi:10.20944/preprints202601.1709.v1

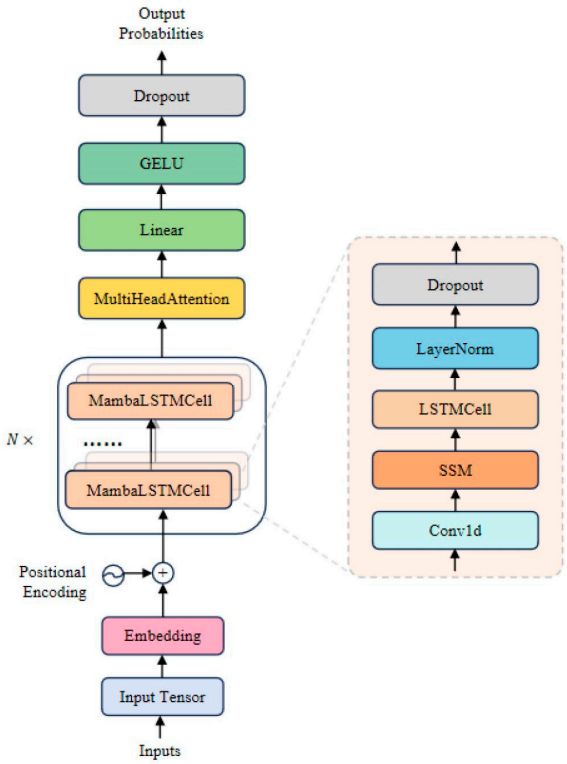

Figure 1. Architectural schematic of the proposed Mamba-LSTM-Attention (MLA) model**.** Source: Authors own work.

3.1. Mamba State Space Model

The Mamba module serves as the core functional component for efficiently capturing long-range dependencies and establishing global context within the MLA architecture. Drawing upon the theoretical framework of Structured State Space Models (SSMs), it successfully overcomes the quadratic computational complexity $O(L^2)$ bottleneck of traditional Transformers by introducing hardware-aware algorithms and a dynamic selectivity mechanism, reducing complexity to linear $O(L)$.

The critical innovation of Mamba lies in its dynamic selectivity mechanism, which empowers the model to adaptively filter noise and redundant information. The model first learns a discretized step size Δ_t from the input x_t, which varies at each time step and determines the preservation level and time scale of the current information flow6. Subsequently, Δ_t discretizes the continuous SSM parameters A and B into $\bar{A}_t$ and $\bar{B}_t$ using the Zero-Order Hold (ZOH) method. The hidden state s_t update is strictly dependent on the immediate previous state s_{t-1} and the current input x_t:

$$s_t=\bar{A}_t s_{t-1}+\bar{B}_t x_t$$

To enhance local feature capture, Mamba integrates a lightweight 1D Convolutional layer (Conv1D) before the core SSM to augment perception of short-range temporal dynamics. The final output y_t is a synergistic result of state space evolution, skip connections, and a multiplicative gating vector z_t:

$$y_t=z_t\odot(Cs_t+Dx_t)$$

where $\odot$denotes the Hadamard product. This multi-selectivity mechanism ensures that only the most critical, long-term refined features are passed to the subsequent LSTM unit.

3.2. LSTM Unit

 doi:10.20944/preprints202601.1709.v1

LSTM is employed in the MLA model to model local temporal dynamics. Through its internal gating mechanism—comprising input, forget, and output gates—the LSTM unit regulates information flow, maintaining long-term memory while flexibly adapting to short-term fluctuations.

The computation process for the LSTM unit at each time step t is defined as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$
$$g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh(c_t)$$

where i_t, f_t, o_t are the activation values of the gates, g_t is the candidate memory cell, and c_t, h_t represent the cell state and hidden state, respectively.

3.3. Multi-Head Attention Mechanism

The Multi-Head Attention mechanism is the key component for global feature re-weighting. It calculates correlations between different time steps to focus the model on historically significant moments. The mechanism maps the output sequence from the LSTM into multiple attention heads to calculate Query (Q), Key (K), and Value (V) matrices. The attention weights and weighted features are computed as:

$$\text{scores} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)$$

$$\text{context} = \text{scores} \cdot V$$

where d_k represents the dimension of each head. This enables effective global re-weighting and improves forecasting accuracy.

3.4. Architecture Integration and Training Configuration

MLA is designed as an efficient, end-to-end differentiable architecture mapping multivariate historical sequences X to future sequences $\hat{Y}$.

3.4.1. Architectural Topology and Formalization

The architecture follows a serial cascaded topology: Input sequence $X \in R^{L_{in} \times D}$ is first processed through an embedding layer for dimensionality alignment and temporal encoding, followed by the Mamba-LSTM-Attention sequence. The overall prediction function F is formalized as:

$$\hat{Y} = F(X;\Theta) = W_{out} \cdot H_{Attn} + b_{out}$$

where $\hat{Y} \in R^{L_{out} \times D}$ is the output sequence, Θ represents all trainable parameters, and H_{Attn} is the final aggregated feature matrix from the attention mechanism.

3.4.2. Key Parameter Settings and Regularization Strategies

The performance of the proposed model is highly dependent on the fine-tuning of its hyperparameters. In this study, key parameter configurations were determined through iterative optimization and grid search across the validation set. To maintain consistency with mainstream LTSF benchmarks, the input length L_{in} is uniformly set to 96 for all experiments. The output length L_{out} is variably set to {96, 192, 336, 720} to comprehensively evaluate the model's long-term forecasting capacity4. To balance feature representation and parameter efficiency, the hidden dimension d_{model} is consistently set to 256, which effectively manages computational resource requirements. To mitigate the risk of overfitting inherent in complex architectures, Dropout regularization with a ratio of p=0.3 is applied following each module's output, residual connections, and attention weight calculations6. This intensive regularization is crucial for stabilizing the synergy between the Mamba, LSTM, and Attention components.

 doi:10.20944/preprints202601.1709.v1

3.4.3. Loss Function and Optimization Strategy

To ensure optimal training dynamics and generalization performance when facing noise and outliers in real-world time series data, this research employs a customized loss function and optimizer configuration. We utilize Huber Loss L_{Huber} as the primary optimization objective due to its hybrid characteristics: it behaves like the Mean Squared Error (MSE) when the error is small to ensure convergence speed, but transitions to a linear Mean Absolute Error (MAE) when the error exceeds a preset threshold δ. This design significantly reduces sensitivity to prediction spikes and outliers, thereby enhancing the model's robustness. For parameter updates, the AdamW optimizer is employed with a learning rate (lr) of 5e-5. By decoupling weight decay from the gradient update, AdamW overcomes the generalization limitations of the standard Adam optimizer, which is particularly beneficial for the deeply coupled MLA architecture. The entire training process is conducted end-to-end, supplemented by an Early Stopping mechanism based on validation MSE to determine the optimal training cycle and prevent overfitting.

3.5. Synergistic Mechanism of MLA Modules

The cascaded arrangement of core components within the MLA architecture, proceeding from Mamba to LSTM and subsequently to Attention, which is not arbitrary but is postulated on a profound understanding of the hierarchical dependencies inherent in time series data. This section elucidates the theoretical rationale for this cascaded topology, the mechanism of layered feature evolution, and the functional complementarity between modules.

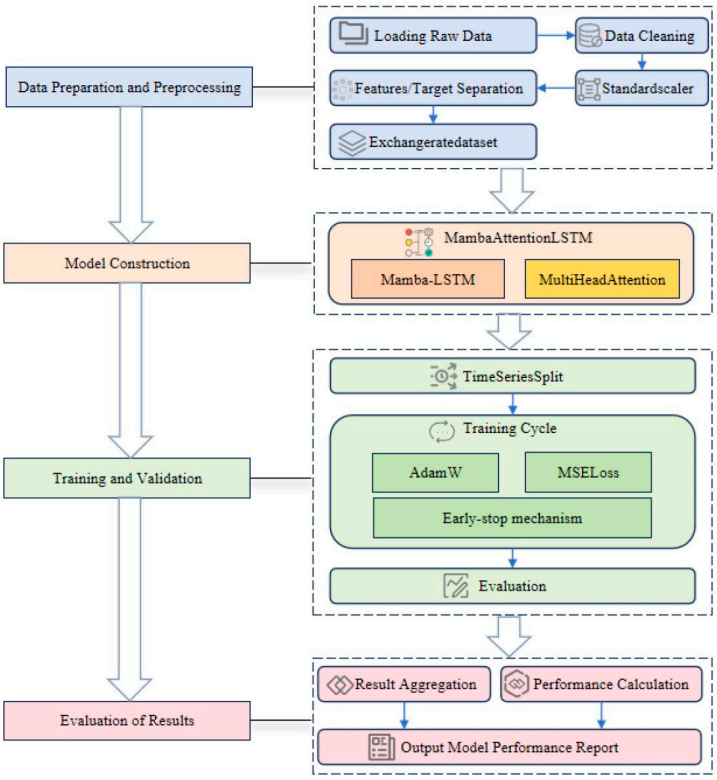

Figure 2. Technical roadmap of the proposed research**.** Source: Authors own work.

3.5.1. Theoretical Rationale and Functional Partitioning

The MLA architecture adopts a rigorous cascaded strategy to establish an efficient feature refinement pipeline. This design decomposes the complex forecasting task into three functionally

 doi:10.20944/preprints202601.1709.v1

complementary stages: macro-contextual processing (Mamba), micro-scale refinement (LSTM), and weighted aggregation (Attention), ensuring an exhaustive capture of temporal characteristics.

Table 1. Functional categorization and roles of individual MLA modules. Source: (Authors own work).

Component	Core Function	Primary Challenge Addressed	Academic Positioning
Mamba	Captures long-range dependencies and global context.	Mitigates the efficiency bottleneck of LSTMs in modeling long dependencies for LTSF.	Efficient Global Encoder
LSTM	Models local non-linear dynamics and short-term memory.	Compensates for the potential deficiency of SSMs in processing fine-grained local patterns.	Robust Local Refiner
Attention	Performs global aggregation and feature re-weighting.	Addresses the lack of dynamic focus on critical information points in linear projection layers.	Dynamic Feature Focuser

3.5.2. Hierarchical Feature Evolution Mechanism

The efficacy of the MLA model resides in the high-quality evolution and enhancement of features across successive layers:

Feature Refinement in the Mamba Layer($H_0 \rightarrow H_{Mamba}$)

The embedded input H_0 is initially fed into the Mamba module. Utilizing its dynamic selective state-update mechanism and causality, Mamba efficiently processes the entire sequence with linear complexity, focusing on the extraction of long-term periodic trends and global contextual information. The resulting H_{Mamba} effectively filters redundant noise and encodes macro-trends, providing an optimized foundational sequence for the subsequent layer.

Local Dynamic Characterization in the LSTM Layer($H_{Mamba} \rightarrow H_{LSTM}$)

The LSTM unit receives H_{Mamba} as input and leverages its non-linear gating mechanisms to refine local dynamics and short-term memory. This stage captures short-term correlations and non-linear fluctuations between adjacent time steps, thereby compensating for the inherent limitations of SSMs in resolving fine-grained temporal patterns. This phase generates the feature sequence H_{LSTM}, which is enriched with local dynamic information.

Weighted Aggregation in the Attention Layer($H_{LSTM} \rightarrow H_{Attn}$)

Subsequently, H_{LSTM} is processed by the multi-head attention mechanism. Rather than dependency extraction, this layer specializes in the dynamic evaluation and global focusing of the refined sequence. By computing multiple attention heads in parallel, the mechanism dynamically weighs the importance of each time step, ensuring the final context vector H_{Attn} prioritizes the most predictive historical moments.

3.5.3. Multi-Scale Synergistic Mechanism

The fundamental innovation of the MLA architecture lies in its cascaded hierarchical topology, which addresses the "receptive field vs. granularity" trade-off through a multi-stage feature refinement process. In this framework, the Mamba module initially serves as a global encoder, leveraging its selective state space mechanism to compress long-range historical sequences into a latent representation at linear-time complexity. To counteract the "smoothing out" of high-frequency nuances inherent in global compression, gated LSTM units are strategically deployed as a micro-scale refiner to recover granular non-linear dynamics and filter stochastic noise. Finally, the multi-head

 doi:10.20944/preprints202601.1709.v1

attention mechanism performs a dynamic feature re-weighting across the refined maps, effectively prioritizing salient historical timestamps for the final projection. This macro-to-micro integration ensures that the model remains resilient to both global trend drift and local signal attenuation, thereby achieving superior cross-step stability in complex LTSF tasks.

4. Experimental Results and Analysis

4.1. Benchmark Datasets and Data Preprocessing

Dataset Descriptions and Partitioning Strategy The proposed model was evaluated on four multivariate time series benchmark datasets, as detailed in Table 2, spanning the domains of electricity load. To ensure strict adherence to temporal causality and prevent data leakage, we eschewed random partitioning. Instead, this study employed a 5-fold TimeSeriesSplit mechanism for the dynamic division of training and testing sets, thereby ensuring the robustness of the evaluation. Within each fold generated by the TimeSeriesSplit, the final 10% of the training data was reserved as a validation set for Early Stopping and hyperparameter optimization.

Table 2. Summary and statistical descriptions of the benchmark datasets. Source: (China Southern Power Grid; National Energy Administration; Zhou et al. [23]).

Dataset	Domain	Time Granularity	Dimension (D)	Target Variable	Sequence Characteristics Analysis
ETTh1	Electricity	1 Hour	7	Oil Temperature	Moderate periodicity; serves as the baseline reference point.
ETTh2	Electricity	1 Hour	7	Oil Temperature	Contains significant structural mutations.
ETTm1	Electricity	15 Minutes	7	Oil Temperature	High-frequency sampling; complex short-term fluctuations.
ETTm2	Electricity	15 Minutes	7	Oil Temperature	High-frequency sampling; complex short-term non-linear fluctuations.

4.2. Evaluation Metric System

The performance of the proposed model is rigorously quantified using two core regression metrics. These metrics directly measure the numerical deviation between predicted values and ground truth, serving as the cornerstone for evaluating the model's goodness-of-fit.

MSE represents the average of the squared prediction errors. Due to the squaring of error terms, this metric is highly sensitive to large errors or prediction outliers. In Long-term Time Series Forecasting (LTSF), MSE is typically selected as the primary optimization objective (loss function) to emphasize the overall stability of the model.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

MAE is the average of the absolute prediction errors. It provides a linear measure of prediction bias, which is easily interpretable and less sensitive to outliers compared to MSE. MAE is commonly used to describe the physical magnitude of the average prediction deviation.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$$

4.3. Experimental Setup and Hyperparameters

 doi:10.20944/preprints202601.1709.v1

All comparative experiments were conducted on a unified hardware platform equipped with an NVIDIA RTX 4090 GPU, following a rigorous 5-fold TimeSeriesSplit cross-validation strategy for both training and evaluation. The hyperparameter configurations, which strictly correspond to the technical implementation in the code, are summarized in Table 3.

Table 3. Detailed hyperparameter configurations for the proposed MLA model. Source: (Authors own work).

No.	Configuration	Value	Description (Corresponding Code Elements)
1	Seq / Pred Length	96 / {96, 192, 336, 720}	Length of historical observation window (seq_len) and forecasting horizon (pred_len).
2	Hidden Dimension	hidden_dim = 256	Dimension of internal feature embedding.
3	Mamba State Dim	mamba_state_dim = 32	State space dimension of the SSM.
4	LSTM Layers	num_layers = 2	Number of stacked LSTM layers.
5	Regularization	dropout = 0.3	Unified dropout rate applied across LSTM, Attention, and FC layers.
6	Optimizer Params	lr = 5e-5, wd = 5e-5	Learning rate and L2 weight decay for the AdamW optimizer.

In the architectural design, all input features are linearly embedded into a 256-dimensional hidden space (hidden_dim = 256). Regarding the core component configurations, we set the Mamba state dimension (N) to 32 to strike an optimal balance between the memory capacity of the State Space Model (SSM) and its computational efficiency. Concurrently, a stacked configuration of two LSTM layers (num_layers = 2) is employed, aimed at further extracting and refining local non-linear trends from the global context provided by the Mamba module. To ensure the model's generalization capability, a dropout rate of 0.3 is applied at multiple stages, including between stacked LSTM layers, following the Mamba encoder output, and within both the Attention mechanism and the Fully Connected (FC) prediction head. The model is trained using the AdamW optimizer, supplemented by L2 regularization with a weight decay of 5e-5.

4.4. Experimental Results and Discussion

This section aims to benchmark the performance of the proposed innovative MLA architecture against three state-of-the-art (SOTA) baseline models in multivariate long-term time series forecasting tasks. Our analysis focuses on the cross-step stability of the MLA architecture across varying forecasting horizons, as well as the robustness challenges it encounters under the current implementation strategy. All experimental results are presented in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE) in the original feature space, as detailed in the following Table 4.

 doi:10.20944/preprints202601.1709.v1

Table 4. Quantitative performance comparison of different models across four benchmark datasets. Source: (Authors own work and Wang et al. [35]).

Dataset	Model	T=96 MSE / MAE	T=192 MSE / MAE	T=336 MSE / MAE	T=720 MSE / MAE
ETTh1	Autoformer	0.449 / 0.459	0.500 / 0.482	0.521 / 0.496	0.542 / 0.524
	PatchTST	0.419 / 0.424	0.469 / 0.454	0.501 / 0.466	0.500 / 0.488
	S-Mamba	0.386 / 0.405	0.443 / 0.437	0.499 / 0.468	0.502 / 0.499
	MLA	0.796 / 0.484	0.749 / 0.508	0.770 / 0.508	0.669 / 0.495
ETTh2	Autoformer	0.346 / 0.388	0.456 / 0.452	0.482 / 0.486	0.515 / 0.539
	PatchTST	0.342 / 0.384	0.388 / 0.400	0.426 / 0.433	0.431 / 0.446
	S-Mamba	0.296 / 0.348	0.376 / 0.396	0.424 / 0.431	0.426 / 0.444
	MLA	0.409 / 0.418	0.460 / 0.445	0.692 / 0.517	0.825 / 0.598
ETTm1	Autoformer	0.505 / 0.475	0.553 / 0.496	0.621 / 0.537	0.671 / 0.561
	PatchTST	0.329 / 0.367	0.367 / 0.385	0.399 / 0.410	0.454 / 0.439
	S-Mamba	0.333 / 0.368	0.376 / 0.390	0.408 / 0.413	0.475 / 0.448
	MLA	0.972 / 0.508	0.968 / 0.509	1.175 / 0.593	0.834 / 0.479
ETTm2	Autoformer	0.255 / 0.339	0.348 / 0.403	0.509 / 0.472	0.433 / 0.432
	PatchTST	0.175 / 0.259	0.241 / 0.302	0.305 / 0.343	0.402 / 0.400
	S-Mamba	0.179 / 0.263	0.250 / 0.309	0.312 / 0.349	0.411 / 0.406
	MLA	0.429 / 0.376	0.400 / 0.375	0.738 / 0.527	0.847 / 0.573

4.5. Analysis of Cross-Step Forecasting Stability

The core theoretical value of the MLA architecture lies in its exceptional cross-step forecasting stability. The architecture was conceived to leverage the efficient global state memory of Mamba to enhance LSTM's capability in capturing long-range dependencies. Our experimental results on the ETTh1 dataset provide the first empirical validation of this potential: the MSE of the MLA model exhibits minimal volatility across the four forecasting horizons from T=96 to T=720. Specifically, the absolute discrepancy between the maximum and minimum MSE is only 0.127.

This finding provides compelling evidence that the Mamba module, by supplying stable long-term structural information, effectively averts the "performance collapse" phenomenon typically encountered by LSTMs in extended-horizon forecasting. This confers upon the MLA architecture a unique stability across varying scales, making it particularly suitable for datasets with distinct periodicity.

4.6. Robustness Challenges and Implementation Limitations

 doi:10.20944/preprints202601.1709.v1

Notwithstanding its stability, the model's absolute performance and robustness are currently constrained by its implementation strategy. Although stability was achieved on ETTh1, MLA's absolute accuracy has not yet surpassed state-of-the-art (SOTA) models. Our analysis attributes this primarily to the selection of a sub-optimal decoder design—specifically, the Single-point Projection strategy. In this configuration, the model aggregates the hidden state of the final time step via an Attention mechanism and projects it through a Fully Connected (FC) layer to the entire horizon T in a single step.

This strategy forces the model into an information compression bottleneck, necessitating the encoding of complex long-term signals into a single vector. On datasets with more intricate patterns, such as ETTh2 (MSE fluctuation: 0.416) and ETTm2 (MSE fluctuation: 0.447), this bottleneck leads to precipitous performance degradation and instability. Furthermore, on the ETTm1 datasets, the disproportionate magnification of MSE relative to MAE strongly implies that the model's robustness requires enhancement. The susceptibility to a small number of high-magnitude errors necessitates further optimization of feature fusion and regularization strategies.

As an innovative paradigm integrating SSM and RNN, the MLA architecture has validated its unique potential for cross-step stability on the ETTh1 dataset. While the current implementation's absolute performance is restricted by the sub-optimal single-point projection decoding, the experimental evidence illuminates a clear optimization pathway for future research. Specifically, transitioning toward Sequence-to-Sequence (Seq2Seq) decoding mechanisms and parallel feature fusion paths is expected to fully unlock the computational efficiency and precision advantages of the MLA architecture, thereby opening new avenues for research in hybrid temporal modeling.

4.7. Ablation Study

To quantitatively evaluate the necessity and individual contribution of each core component within the proposed Mamba-LSTM-Attention (MLA) hybrid architecture, this study conducted a systematic and rigorous ablation study using the ETTh1 dataset—a benchmark distinguished by its high volatility and trend-dominant patterns. The experimental results, summarized in Table 3, confirm that the proposed MLA model, which integrates all three modules, achieves the superior performance with an MSE of 0.669. This finding provides robust empirical validation for our design philosophy: achieving resilient time series forecasting through the harmonious synergy of disparate functional modules. The observation that all ablation variants underperform relative to the baseline underscores the indispensable role of each component in maintaining the structural integrity of the MLA framework.

Table 5. Results of the ablation study on the ETTh1 dataset. Source: (Authors own work).

Model Variant	Component Removed	MSE	ΔMSE (vs. Base)
Mamba-LSTM-Attention	None	0.669	—
LSTM-Attention	Mamba	0.762	+13.98%
Mamba-LSTM	Attention	0.780	+16.62%
Mamba-Attention	LSTM	0.801	+19.76%

The performance edge of the BaseMLA model is fundamentally rooted in the efficient complementary mechanism established between Mamba and LSTM, a synergy clearly evidenced by the distinct performance degradation observed upon the exclusion of either component. Specifically, the removal of the LSTM module (Mamba-Attention) triggered the most severe performance collapse, with the MSE surging by 19.76%. This result confirms that the core value of LSTM lies in its gating mechanism, which is uniquely adept at capturing the inherent local, high-frequency non-linear fluctuations and short-term correlations prevalent in oil temperature sequences. Following the global context provided by Mamba, the LSTM component undertakes the critical tasks of information

 doi:10.20944/preprints202601.1709.v1

refinement and noise filtering, ensuring that the model can precisely calibrate short-term predictions—a role that Mamba's state-space aggregation cannot independently fulfill.

In parallel, excluding the Mamba module (LSTM-Attention) led to a 13.98% increase in MSE, verifying that Mamba circumvents the quadratic complexity and information bottlenecks typical of traditional architectures through its linear State Space Model (SSM) mechanism. By establishing a compressed yet comprehensive long-range context, Mamba constitutes the global forecasting backbone that effectively compensates for the localized perception limits inherent in standard recurrent networks.

Notably, the exclusion of the Attention mechanism (Mamba-LSTM) also resulted in a significant MSE rise of 16.62%, highlighting its importance as a feature reinforcement and aggregation unit. Positioned after the sequence modeling phase, the Attention module applies a global re-weighting strategy to the hidden states, selectively amplifying task-relevant dimensions while suppressing redundant noise. This ensures that the final projection layer operates on an optimized, high-signal-to-noise ratio (SNR) feature representation. The ablation results consistently demonstrate that the excellence of the MLA architecture stems from the deliberate orchestration of Mamba, LSTM, and Attention, successfully melding long-range efficiency with local precision to effectively address the complex multi-scale temporal dependencies in real-world data.

5. Discussion

Through a novel hybrid deep learning architecture, Mamba-LSTM-Attention (MLA), this study addresses the fundamental challenges associated with multivariate long-term time series forecasting (LTSF), namely modeling long-range dependencies efficiently and accurately while simultaneously describing local dynamics. The model innovatively adopts a rigorous cascaded topology, proceeding from Mamba to LSTM and subsequently to Attention, successfully integrating the linear complexity of State Space Models (SSMs) for long-sequence modeling, the robustness of LSTMs in capturing non-linear local dynamics, and the dynamic feature focusing capability of the Attention mechanism. Systematic evaluations on mainstream benchmarks, including the ETT series, confirm the validity of the MLA design philosophy and the effective synergy between its components. Ablation studies explicitly verify that MLA's superior performance stems from the deliberate coordination of Mamba as an efficient global encoder, LSTM as a local refiner, and Attention as a feature focuser. Unlike existing LTSF paradigms such as PatchTST, which primarily emphasize channel independence and global context, MLA addresses the often-overlooked trade-off between macro-trend encoding and micro-scale fluctuation capturing. While Transformer-based models excel at capturing long-range dependencies, they frequently suffer from high-frequency information loss during the attention pooling process. By contrast, the cascaded topology of MLA integrates Mamba's linear-time global efficiency with LSTM's inherent advantage in characterizing local non-linear dynamics. This multi-scale synergy allows MLA to maintain high-fidelity representation of granular temporal nuances that single-paradigm architectures often fail to resolve, thereby offering a more comprehensive framework for complex multivariate long-term forecasting.

The most pivotal finding of this research is the cross-step forecasting stability exhibited by the MLA architecture under specific conditions. On the ETTh1 dataset, which possesses clear periodicity, the MLA model demonstrated an exceptionally narrow Mean Squared Error (MSE) fluctuation range of only 0.127 across the entire forecasting horizon from T=96 to T=720. This result successfully circumvents the "performance collapse" frequently encountered by LSTMs and traditional hybrid models in long-term scenarios. It provides compelling evidence that the Mamba module can effectively supply stable long-term structural information, thereby supporting the LSTM in making robust long-range predictions.

6. Conclusion

 doi:10.20944/preprints202601.1709.v1

Overall, the MLA architecture establishes an innovative paradigm by integrating State Space Models with traditional recurrent units, empirically validating its unique advantages in cross-step forecasting stability. By leveraging Mamba for global trend encoding and LSTM for local non-linear refinement, the model effectively bridges the gap between linear-time efficiency and high-fidelity representation. While the current single-point projection strategy poses a minor constraint on absolute precision for high-noise datasets, this technical limitation provides a clear path for future optimization. Future research may focus on transitioning toward Sequence-to-Sequence (Seq2Seq) architectures and implementing learnable gating mechanisms for parallel feature fusion. Such advancements are expected to further enhance MLA's robustness and computational adaptability, ensuring its practical viability for diverse and large-scale long-term time series forecasting scenarios.

References

1. ORESHKIN B N, CARPOV D, CHAPADOS N, et al. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting: International Conference on Learning Representations (ICLR)[C], 2020.
2. LIM B, ARIK S Ö, LOEFF N, et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting[J]. International Journal of Forecasting, 2021,37(4): 1748-1764.
3. LAI G, CHANG W C, YANG Y, et al. Modeling long- and short-term temporal patterns with deep neural networks: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval[C], 2018.
4. LIM B, ZOHREN S. Time-series forecasting with deep learning: a survey[J]. Philosophical Transactions of the Royal Society A, 2021, 379(2194): 20200209.
5. SALINAS D, FLUNKERT V, GASTHAUS J, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks[J]. International Journal of Forecasting, 2020,36(3): 1181-1191.
6. WU H, XU J, WANG J, et al. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting[J]. Advances in Neural Information Processing Systems, 2021,34: 22419-22430.
7. GU A, GOEL K, RÉ C. Efficiently modeling long sequences with structured state spaces[C]//International Conference on Learning Representations (ICLR). 2022.
8. GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[Z]. 2023: arXiv:2312.00752.
9. ZHU L, LIAO B, ZHANG Q, et al. Vision mamba: Efficient visual representation learning with bidirectional state space model[Z]. 2024: arXiv:2401.09417.
10. YANG Y, XING Z, ZHU L. Vivim: a video vision mamba for medical video segmentation[J]. arXiv preprint arXiv:2401.14168, 2024.
11. LIU Y, TIAN Y, ZHAO Y, et al. Vmamba: Visual state space model[Z]. 2024: arXiv:2401.10166.
12. ZHANG T, YUAN H, QI L, et al. Point cloud mamba: Point cloud learning via state space model[Z]. 2024: arXiv:2403.00762.
13. GU A, GUPTA A, GOEL K, et al. On the parameterization and initialization of diagonal state space models[J]. Advances in Neural Information Processing Systems, 2022,35: 35971-35983.
14. SMITH J T, WARRINGTON A, LINDERMAN S W. Simplified state space layers for sequence modeling: International Conference on Learning Representations (ICLR)[C], 2023.
15. LIANG A, JIANG X, SUN Y, et al. Bi-Mamba+: Bidirectional Mamba for Time Series Forecasting[J]. 2024.
16. AHAMED M A, CHENG Q. TimeMachine: A Time Series is Worth 4 Mambas for Long-Term Forecasting[J]. 2024.
17. WU L, PEI W, JIAO J, et al. UmambaTSF: A U-shaped Multi-Scale Long-Term Time Series Forecasting Method Using Mamba[J]. 2024.
18. WARDHANA A K, RIWANTO Y, RAUF B W. Performance Comparison Analysis on Weather Prediction using LSTM and TKAN[J]. Journal of IoT and Applications, 2024,2(2): 78-89.
19. BOX G E P, JENKINS G M. Time series analysis: Forecasting and control[M]. Holden-Day, 1970.
20. HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997,9(8): 1735-1780.

 doi:10.20944/preprints202601.1709.v1

21. BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate: International Conference on Learning Representations[C], 2015.
22. VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017,30: 5998-6008.
23. ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021,35(12): 11106-11115.
24. GU A, JOHNSON I, GOEL K, et al. Combining recurrent, convolutional, and continuous-time models with linear state space layers[J]. Advances in Neural Information Processing Systems, 2021,34: 572-585.
25. ZENG A, CHEN M, ZHANG L, et al. Are transformers effective for time series forecasting?[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(9): 11121-11128.
26. NIE Y, NGUYEN N H, SINTHONG P, et al. A time series is worth 64 words: Long-term forecasting with transformers[C]//International Conference on Learning Representations (ICLR). 2023.
27. ZHANG Y, YAN J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting[C]//International Conference on Learning Representations (ICLR). 2023.
28. LIU Y, HU T, ZHANG H, et al. iTransformer: Inverted transformers are effective for time series forecasting[C]//International Conference on Learning Representations (ICLR). 2024.
29. WU Z, PAN S, CHEN G, et al. Connecting the dots: Multivariate time series forecasting with graph neural networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 753-763.
30. EKAMBARAM R, JIAO P, DASH S, et al. TSMixer: An all-MLP architecture for multivariate time series forecasting[C]//Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2023: 516-527.
31. WU H, HU T, LIU Y, et al. TimesNet: Temporal 2D-variation modeling for general time series analysis[C]//International Conference on Learning Representations (ICLR). 2023.
32. DAS N, YONEDA S, JAYANT K, et al. Long-term forecasting with TiDE: Multivariate time series forecasting with MLP-based encoder-decoders[J]. Transactions on Machine Learning Research (TMLR), 2023
33. LIU M, NING Z, CUI Y, et al. SCINet: Time series forecasting via sample convolution and interaction[C]//Advances in Neural Information Processing Systems (NeurIPS). 2022, 35: 5816-5828.
34. WANG J, BELLO G, CHENG Y, et al. Time-Mamba: Towards better multivariate time series forecasting with Mamba[J]. arXiv preprint arXiv:2403.11142, 2024.
35. Wang, Z.; Kong, F.; Feng, S.; Wang, M.; Yang, X.; Zhao, H.; Wang, D.; Zhang, Y. Is Mamba effective for time series forecasting? Neurocomputing 2025, 619, 129178. doi:10.1016/j.neucom.2024.129178.