Transformer Architectures for Time Series Analysis: "A Survey of Methods, Challenges, and Future Directions

[1]**Kavita Kumari**
[2]**Yazdani Hasan**
Yudesh.2008@gmail.com
yazhassid@gmail.com

Abstract —Transformer architectures have revolutionized deep learning across natural language processing and computer vision, and have recently emerged as the dominant paradigm for time series analysis. This paper provides a comprehensive survey of transformer-based methods for time series forecasting, classification, anomaly detection, and representation learning. We systematically review architectural innovations including efficient attention mechanisms, channel-independent designs, and adaptations for multi-scale temporal dependencies. Drawing on a systematic analysis of recent literature, we present a task-aware taxonomy that organizes transformer variants according to their architectural design and application domain. We identify persistent challenges including computational scalability with long sequences, data efficiency constraints, overfitting risks, and interpretability limitations. Finally, we outline promising future directions encompassing physics-informed architectures, resource-efficient inference, multi-resolution spatiotemporal learning, and integration with foundation models. This survey aims to provide researchers and practitioners with a structured foundation for advancing transformer-based time series analysis.

Keywords:Transformers, Time Series Analysis, Attention Mechanisms, Sequence Modeling, Deep Learning

1. Introduction

Time series data underpin critical decision-making processes across finance, healthcare, energy systems, industrial monitoring, and climate science. Accurate forecasting, anomaly detection, and representation learning from temporal data enable applications ranging from predictive maintenance to epidemic tracking and algorithmic trading .

For decades, time series analysis was dominated by statistical approaches such as ARIMA and exponential smoothing, later augmented by recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. However, these architectures exhibit fundamental limitations: RNNs struggle with long-range dependencies due to vanishing gradients, while convolutional neural networks require deep stacking to achieve adequate receptive fields .

The introduction of the Transformer architecture by Vaswani et al. (2017) marked a paradigm shift. Its core innovation—the self-attention mechanism—enables direct modeling of relationships between all time steps regardless of distance, providing flexible temporal receptive fields and parallelizable training. Since 2020, transformer-based models have achieved state-of-the-art performance across virtually all time series benchmarks .

This paper provides a systematic survey of transformer architectures for time series analysis. We address three research questions: (1) What architectural innovations have been developed to adapt transformers for temporal data? (2) How do these methods perform across different time series tasks? (3) What challenges remain, and what future directions are most promising?

The remainder of this paper is organized as follows. Section 2 presents foundational concepts in time series analysis and transformer architectures. Section 3 provides a task-aware taxonomy of transformer methods. Section 4 examines key architectural innovations. Section 5 discusses evaluation benchmarks and performance comparisons. Section 6 identifies persistent challenges, and Section 7 outlines future research directions. Section 8 concludes.

2. Foundations of Time Series Analysis with Transformers

2.1 Characteristics of Time Series Data

Time series data exhibit several distinctive properties that influence architectural design :

Temporal dependencies range from short-term patterns (hourly cycles) to long-range trends (yearly seasonality). Effective models must capture dependencies across multiple scales simultaneously.

Non-stationarity means statistical properties change over time due to regime shifts, concept drift, or external interventions. Models must adapt to distributional shifts.

Multi-resolution structure arises from hierarchical temporal patterns—minutes within hours within days within seasons.

Multivariate interactions in systems like power grids or financial markets involve complex cross-series dependencies that must be modeled alongside temporal dynamics.

2.2 The Transformer Architecture

The canonical Transformer processes sequences through stacked encoder layers, each containing multi-head self-attention (MHSA) and feed-forward networks with residual connections and layer normalization . For time series, the input sequence $X \in \mathbb{R}^{L \times d}$ (length L, d features) is projected to queries Q, keys K, and values V:

$$
\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
$$

Multi-head attention runs this mechanism in parallel across h subspaces, enabling the model to attend to information from different representation dimensions.

The position encoding mechanism—critical for sequence order—has been substantially rethought for time series, with learned embeddings, temporal windows, and timestamp-based encodings replacing original sinusoidal versions .

3. A Task-Aware Taxonomy of Transformer Methods

Transformers have been adapted for diverse time series tasks through specialized architectural designs .

3.1 Forecasting Models

Time series forecasting remains the most active application area. Transformer-based forecasters generally follow encoder-only, decoder-only, or encoder-decoder architectures.

Encoder-only models like Autoformer and FEDformer process historical windows and project to future predictions through linear heads. These models emphasize efficient attention mechanisms to handle long sequences.

Decoder-only architectures (e.g., Transformer decoder variants) generate forecasts autoregressively, conditioning on previous predictions. While flexible, they risk error accumulation over long horizons.

Encoder-decoder models such as Informer and Pyraformer separate context encoding from prediction generation, often incorporating specialized attention mechanisms for multi-scale processing.

Recent innovations include channel-independent architectures that process each time series variate separately before fusion, reducing computational complexity while maintaining accuracy .

3.2 Anomaly Detection

Transformer-based anomaly detection leverages the reconstruction error or attention patterns to identify unusual temporal patterns. Models like TranAD and Anomaly Transformer use association discrepancy between adjacent time steps to highlight anomalous segments .

3.3 Representation Learning

Self-supervised pre-training of transformers on large time series corpora has emerged as a promising direction. Models learn general-purpose temporal representations through masked reconstruction, contrastive learning, or forecasting objectives, then fine-tune for downstream tasks .

3.4 Multimodal Fusion

Time series often coexist with other modalities—text reports in healthcare, images in industrial inspection, or sensor metadata. Cross-modal transformers enable joint representation learning across heterogeneous data types .

4. Key Architectural Innovations

Adapting transformers to time series has required fundamental architectural innovations addressing efficiency, scalability, and temporal structure.

4.1 Efficient Attention Mechanisms

The quadratic complexity \(O(L^2)\) of full attention prohibits application to long sequences. Several efficient variants have emerged :

Sparse attention patterns restrict each token's attention field to local windows or random subsets. LogSparse attention and Longformer-style sliding windows reduce complexity to \(O(L \log L)\).

ProbSparse attention (in Informer) identifies dominant query-key pairs through KL divergence measurement, computing attention only for the most informative combinations.

Auto-correlation mechanisms (in Autoformer) replace dot-product attention with series-wise autocorrelation, capturing period-based dependencies with \(O(L \log L)\) complexity.

Linear attention reformulates attention with kernel methods, achieving linear complexity while maintaining reasonable expressivity.

4.2 Multi-Scale Processing

Time series exhibit patterns across multiple temporal scales. Architectures address this through :

Hierarchical designs like Pyraformer construct pyramid attention graphs where fine-scale tokens attend to coarser-scale representations.

Dilated convolution integration (as in MDCSformer) combines multi-scale dilated convolutions for local feature extraction with attention for global context .

Multi-resolution branching processes time series at different temporal resolutions in parallel, fusing representations adaptively.

4.3 Position and Temporal Encoding

Beyond absolute position, time series require encoding of timestamps, periods, and intervals. Modern approaches include learnable temporal embeddings, Fourier features, and relative position biases that capture temporal distances .

4.4 Lightweight Architectures for Edge Deployment

Recent work has focused on compressing transformers for resource-constrained environments. Techniques include knowledge distillation (DistilBERT-style), quantization to INT8/FP16, pruning, and hardware-aware neural architecture search. Modern lightweight transformers achieve 75-96% of full-model accuracy while reducing size by 4-10× and latency by 3-9×, enabling deployment on devices with 2-5W power consumption .

5. Evaluation Benchmarks and Performance

5.1 Standard Datasets

Transformer evaluation spans multiple benchmark families :

Monash Forecasting Repository includes over 30 datasets spanning energy, finance, healthcare, and tourism with varying frequencies and lengths.

ETT (Electricity Transformer Temperature) provides multi-variate time series from electricity transformers, widely used for long-sequence forecasting.

Weather, Traffic, and Exchange Rate datasets from Informer and Autoformer papers enable standardized comparisons.

ImageNet-1K and COCO adaptations for time series-to-vision tasks evaluate spatiotemporal models.

5.2 Key Findings

Empirical studies reveal several consistent patterns :

- Efficient attention mechanisms achieve comparable accuracy to full attention at substantially lower computational cost.
- Channel independence often improves multivariate forecasting by avoiding spurious cross-channel correlations.
- Model size sweet spots : 15-40M parameter models achieve optimal hardware utilization (60-75% efficiency) on edge devices.
- Quantization trade-offs : INT8 quantization preserves accuracy for most models, while FP16 offers better gradient stability for training.
- Hybrid CNN-Transformer architectures frequently outperform pure transformers on noisy or short-sequence data by combining local feature extraction with global context .

6. Persistent Challenges

Despite remarkable progress, significant challenges remain .

6.1 Computational Scalability

Even with efficient attention, processing extremely long sequences (e.g., years of minutely data) remains prohibitive. Memory constraints limit context windows, forcing trade-offs between historical depth and model complexity.

6.2 Data Efficiency

Transformers are notoriously data-hungry, requiring extensive training data to outperform simpler methods. In small-sample regimes (common in healthcare and industrial applications), they often underperform tuned statistical models or regularized linear methods .

6.3 Overfitting and Generalization

The high capacity of transformers creates overfitting risks, particularly for noisy time series. Distribution shift between training and deployment further degrades performance—a model trained on historical patterns may fail when regime changes occur.

6.4 Interpretability

While attention weights are often presented as explanations, their relationship to model decisions remains poorly understood. Attention maps may highlight spurious correlations rather than causal structures, limiting trust in high-stakes applications.

6.5 Hyperparameter Sensitivity

Transformer performance depends critically on architecture choices (heads, layers, dimensions) and training hyperparameters. Optimal configurations vary substantially across datasets, necessitating expensive search.

6.6 Reproducibility

Variations in implementation, data preprocessing, and evaluation protocols hinder fair comparison. Recent surveys call for standardized benchmarking frameworks .

7. Future Directions

7.1 Physics-Informed Architectures

Incorporating domain knowledge through physics-informed neural networks represents a promising direction. By embedding differential equations or physical constraints into transformer layers, models can achieve better generalization with less data .

7.2 Hybrid Neural-Mechanistic Modeling

Combining transformers with mechanistic models (e.g., epidemiological compartments, economic equilibrium models) enables interpretable forecasting that respects domain constraints while learning complex patterns from data.

7.3 Resource-Efficient Real-Time Inference

For edge deployment, research is needed on adaptive computation—dynamically adjusting model depth or attention span based on input complexity or latency requirements. Hardware-aware architecture search tailored to specific deployment targets (NVIDIA Jetson, Apple Neural Engine, ARM) can further optimize efficiency .

7.4 Multi-Resolution Spatiotemporal Learning

Real-world systems exhibit both temporal dynamics and spatial dependencies (e.g., traffic networks, power grids). Architectures that jointly model spatiotemporal structure through graph transformers or mesh attention are under active development .

7.5 Foundation Models for Time Series

Inspired by large language models, researchers are exploring pre-trained foundation models for time series. These models, trained on diverse temporal corpora, could enable few-shot adaptation to new tasks and domains. Early prototypes (e.g., TimesFM, Lag-Llama) show promise but require orders-of-magnitude scaling .

7.6 Human-AI Collaborative Paradigms

As time series analysis supports critical decisions, frameworks that enable effective human-AI collaboration are essential. This includes uncertainty estimation, interactive explanation interfaces, and mechanisms for incorporating human feedback during deployment .

8. Conclusion

Transformer architectures have fundamentally transformed time series analysis, achieving state-of-the-art performance across forecasting, anomaly detection, and representation learning tasks. This survey has provided a structured overview of architectural innovations, task-specific adaptations, evaluation benchmarks, and persistent challenges.

Key takeaways include: (1) Efficient attention mechanisms and multi-scale processing are essential for handling temporal data characteristics; (2) Performance gains depend critically on dataset scale and task type—transformers excel with large, complex datasets but may underperform simpler methods in small-sample regimes; (3) Significant challenges remain in computational scalability, interpretability, and generalization under distribution shift; (4) Future directions including physics-informed architectures, foundation models, and human-AI collaboration offer promising paths forward.

As time series data continue to grow in volume and importance across domains, transformer-based methods will remain at the forefront of research and application. By consolidating current knowledge and identifying open problems, this survey aims to support ongoing advances in this rapidly evolving field.

References

[1] A. Vaswani et al., "Attention is All You Need," NeurIPS , 2017.

[2] H. H. Samson, "Lightweight Transformer Architectures for Edge Devices in Real-Time Applications," arXiv preprint arXiv:2601.03290 , 2026.

[3] J. Zhao et al., "A Survey of Transformer Networks for Time Series Forecasting," Computer Science Review , vol. 60, 2026.

[4] C. H. Park, "MDCSformer: Multi-scale dilated contracted sub-transformer for fault diagnosis of rotating machinery," Journal of Computational Design and Engineering , vol. 13, no. 1, pp. 462-485, 2026.

[5] H. Zhou et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," AAAI , 2021.

[6] T. Brown et al., "Language Models are Few-Shot Learners," NeurIPS , 2020.

[7] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL , 2019.

[8] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR , 2021.

[9] S. Zheng et al., "Rethinking Positional Encoding in Transformers for Time Series Forecasting," ICLR , 2024.

[10] Y. Liu et al., "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," NeurIPS , 2021.

[11] S. Li et al., "Enhancing Locality and Long-range Dependencies in Transformers for Time Series Forecasting," NeurIPS , 2023.

[12] T. Chen et al., "TVM: An Automated End-to-End Optimizing Compiler for Deep Learning," OSDI , 2018.

[13] G. Lai et al., "Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks," SIGIR , 2018.

[14] B. Lim et al., "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," International Journal of Forecasting , 2021.

[15] Q. Wen et al., "Transformers in Time Series: A Survey," arXiv preprint arXiv:2202.07125 , 2023.

[16] S. Tang et al., "Time Series Forecasting with Transformer Architectures: A Survey and Empirical Study," IEEE Transactions on Pattern Analysis and Machine Intelligence , 2025.

[17] C. Zhang et al., "A Survey on Transformer Compression," arXiv preprint arXiv:2402.05964 , 2024.

[18] R. Kidger et al., "Neural Controlled Differential Equations for Irregular Time Series," NeurIPS , 2020.