Proceedings of CONF-MSS 2025 Symposium: Machine Vision System
DOI: 10.54254/2755-2721/2025.CH24699

Transformer-KAN: Attention-Guided Kolmogorov-Arnold Networks for Robust Seismic Signal Classification

Ruoxi Chen

Hainan International College, Minzu University of China, Lingshui, China
crx.approaching@gmail.com

Abstract. This paper introduces Transformer-KAN, an innovative model that combines the attention mechanism of Transformers with the Kolmogorov-Arnold Networks (KAN) to achieve robust seismic signal classification. Leveraging a hierarchical attention mechanism to capture both local details and global patterns of seismic waveforms, and employing learnable functional mappings in the KAN-Former module to replace conventional feedforward layers, the model achieves notable improvements in classification accuracy and noise robustness over traditional convolutional networks and standalone Transformer models. Additionally, the incorporation of hybrid positional encoding enhances the model's adaptability to the temporal features of seismic signals. These design innovations enable the Transformer-KAN architecture to effectively extract long- and short-term features and model nonlinear relationships, ensuring stable performance in complex environments. This establishes a reliable foundation for seismic early warning systems.

Keywords: Transformer-KAN, Seismic Signal Classification, Hierarchical Attention Mechanism, Hybrid Position Encoding

1. Introduction

Accurate classification of seismic signals and noise is a pivotal component of seismic early warning systems (EEWs), playing a critical role in reducing false alarm rates and enhancing response efficiency. Seismic records are non-stationary signals [1]; in complex geological settings, noise interference (e.g., anthropogenic activities, instrumental drift, or environmental vibrations) significantly reduces the precision of seismic signal identification [2], thereby impairing disaster response efficacy. The challenge lies in distinguishing between natural seismic events and anthropogenic explosions based on waveform features—an imperative task for seismic research and its applications. Thus, the development of robust classification algorithms is crucial for reliable seismic warnings.

Traditional seismic signal classification methods primarily rely on time-frequency analysis [3] and statistical features (e.g., peak amplitude, signal energy), often combined with threshold- based decisions or template matching. However, these approaches struggle with nonlinear waveform features and low signal-to-noise ratio (SNR) scenarios. Recent machine learning methods, such as support vector machines and random forests, improve classification performance through automated

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.CH24699

feature learning but suffer from limited representational capacity for large-scale data and high computational costs.

Deep learning methods, including convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have demonstrated substantial improvements in seismic signal classification accuracy [4]. For instance, Transformer architectures utilize attention mechanisms to capture long-range waveform dependencies, while lightweight models (e.g., MobileNet) optimize real-time performance. However, current models face challenges, such as excessive parameterization, limited interpretability, and inadequate robustness to local burst noise.

To address these issues, we propose a hybrid architecture—Transformer-KAN—that integrates hierarchical attention mechanisms, KAN modules, and hybrid positional encoding. This design enhances noise suppression capabilities while maintaining real-time performance, offering a reliable classification framework for seismic early warning systems.

2. Related work

2.1. Transformer models and attention mechanisms

The Transformer model, first introduced by Vaswani et al. for natural language processing, is centered on the self-attention mechanism, which efficiently models dependencies between any positions within a sequence [5]. Self-attention facilitates information interaction by computing the similarity between queries, keys, and values (Q, K, V):

$$Attention\,(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}} + M_{seismic})V \tag{1}$$

Here, d_k represents a scaling factor, QK^T computes the dot-product correlation between queries and keys ,and the softmax function calculates attention weights for a weighted sum. In practice, the Transformer employs multi-head attention to parallelize multiple attention heads, focusing on diverse subspaces. The outputs are concatenated and transformed linearly to derive the final representation:

$$MultiHead\,(Q,K,V) = |head_1,\ldots,head_h|W^0\ ,\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

Multi-head attention enhances the model's ability to focus on varying features. In Transformer encoders, the self-attention layer is followed by a feedforward neural network, stabilized by residual

connections and layer normalization. Without convolution or recurrent structures, Transformers can process long sequences in parallel, achieving remarkable success in language and vision domains.

In seismic signal processing, Transformers also exhibit great potential. Their global attention mechanism captures long-term phase dependencies, facilitating the distinction between event signals and background noise [6]. For instance, Mousavi et al.'s "Earthquake Transformer" incorporates hierarchical attention to combine full-waveform and phase-level information, outperforming traditional methods. These advancements demonstrate the efficacy of attention mechanisms in improving seismic signal precision, particularly when balancing local waveform details and global patterns. This foundation underpins our research.

DOI: 10.54254/2755-2721/2025.CH24699

2.2. Kolmogorov-Arnold Networks (KAN)

Kolmogorov-Arnold Networks (KAN) draw inspiration from the Kolmogorov-Arnold representation theorem, which posits that any -dimensional continuous function can be expressed as a composition of univariate continuous functions. Specifically, at most univariate continuous functions suffice to precisely represent an -dimensional function. This theorem provides a theoretical basis for high-dimensional function approximation [7].

KAN leverages this principle by introducing learnable univariate function modules in place of fixed linear weights, enhancing the model's capability to represent complex multivariable functions. [8] Unlike traditional multilayer perceptrons (MLPs), where nodes apply fixed nonlinear activations (e.g., ReLU, Sigmoid), KAN assigns trainable univariate functions to connections. Consequently, KAN tailors nonlinear relationships to specific needs. For instance, the computation at a hidden layer node $X = (x_1, \ldots, x_n)$ can be expressed as:

$$h_j = \sum_{i=1}^{n} g_{ij}(x_i),\ j = 1, 2,\ \ldots, m, \tag{3}$$

Here, g_{ij} denotes the learnable function mapping input to hidden node . The output layer further processes hidden layer outputs with additional univariate functions, yielding the final result:

$$y = \sum_{j=1}^{m} f_j(h_j) = \sum_{j=1}^{m} f_j\left(\sum_{i=1}^{n} g_{ij}(x_i)\right) \tag{4}$$

KAN's adaptability in approximating nonlinear relationships makes it highly suitable for seismic signals characterized by complex waveform patterns. In this study, we integrate KAN into the Transformer framework to improve its representation of nonlinear seismic features.

3. Proposed method

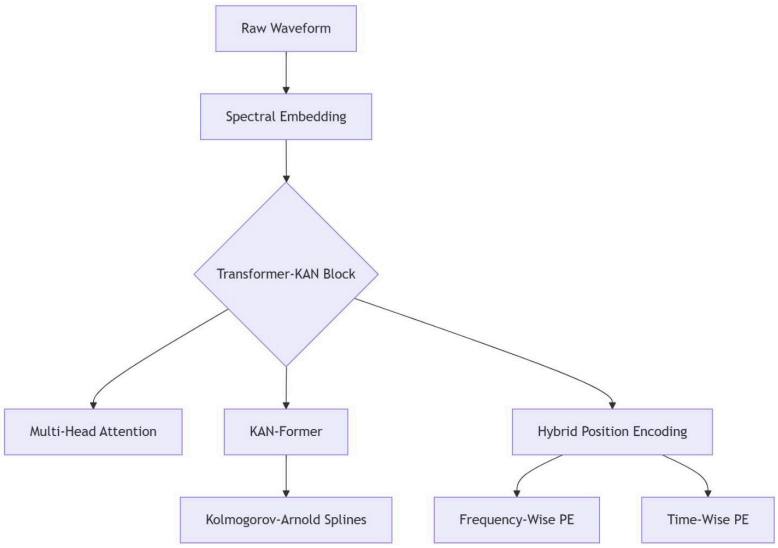

Figure 1. Flowchart

In this research, we propose an innovative Transformer-KAN (KAN-Former) hybrid architecture specifically designed for seismic signal classification tasks. This method organically integrates multi- head self-attention mechanisms with Kolmogorov-Arnold Networks (KAN). Through the

DOI: 10.54254/2755-2721/2025.CH24699

introduction of hierarchical attention mechanisms, KAN modules, and hybrid positional encoding, the model aims to enhance performance, interpretability, and computational efficiency. The following sections detail the proposed model architecture and its components.

3.1. Hierarchical attention mechanism

The Global Attention Mechanism (GAM) constitutes an integral component of the Transformer encoder, designed to process the entire input sequence. Conventional Transformer models employ a singular attention mechanism to capture global dependencies within signals. However, in seismic signal processing, both local and global features exhibit differential significance for classification outcomes. Accordingly, we have developed a hierarchical attention mechanism capable of simultaneously capturing both global and local characteristics of signals.

3.1.1. Global attention layer

The global attention layer extracts global features from the entire input signal to improve the model's understanding of seismic patterns. Its calculation is given as:

$$Attention\,(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}} + M_{seismic})V$$

Here, Q, K, and V represent query, key, and value, respectively. $\frac{QK^T}{\sqrt{d_k}}$ represents the standard self-attention calculation. To incorporate prior knowledge of seismic signals, we introduce a seismic mask matrix $Mseismic$, which integrates a priori information about fault propagation. During global attention computation, the model can focus on physically meaningful regions based on this mask matrix. For instance, in areas with known fault lines, the mask matrix adjusts attention weights to ensure that the model pays more attention to signal patterns related to seismic activities induced by faults.

3.1.2. Local attention window

Local features in seismic signals, such as P-wave arrivals and S-wave conversions, are critical for classification tasks. To this end, we designed a local attention mechanism that employs a 30-second sliding window to capture local time-domain features within the signal. The local attention window facilitates the capture of short-term signal variations, particularly in critical regions of P-to-S wave transitions. By adding this local window mechanism on top of global attention, the model can effectively extract more refined temporal features. For example, during the P-to-S wave conversion phase— which typically contains key information for classifying earthquake types and magnitudes —the local attention window can acutely capture these subtle changes.

3.2. Kan-former module

KANs represent a novel neural network architecture that enhances network performance and interpretability by replacing weight parameters with learnable univariate functions, enabling seamless interaction with human users. To strengthen the model's capability in handling complex nonlinear features within seismic signals, we propose an innovative design of integrating the Kolmogorov- Arnold Network (KAN) into the Transformer architecture. The KAN network, a deep learning model grounded in the Kolmogorov-Arnold representation theorem, is capable of learning

DOI: 10.54254/2755-2721/2025.CH24699

complex signal patterns in high-dimensional spaces through nonlinear transformations. In this model, we substitute the feedforward network (FFN) in the traditional Transformer with a KAN module to enhance the representation ability for complex seismic waveforms.

The core formula of the KAN module is as follows:

$$KAN(x) = \sum_{i=1}^{n} \Phi_i(\sum_{j=1}^{m} \psi_{ij}(x_j)) \tag{5}$$

where Φ_i denotes the basis function and ψ_{ij} is the B-spline kernel function. The B-spline kernel exhibits strong local approximation capability, enabling effective processing of nonlinear features in seismic signals. By replacing the traditional components with the KAN network, the model not only achieves better fitting of complex seismic signal patterns but also enhances its nonlinear mapping ability, thereby improving classification accuracy.

The advantage of the KAN network lies in its nonlinear fitting capability for data through basis functions, which allows it to adapt to the highly nonlinear and complex waveform variations in seismic signals. Specifically, KAN maps input signals to high-dimensional spaces, learning the nonlinear laws of seismic waveforms to enhance signal representation. After replacing the traditional FFN, the model can more effectively capture the complex features of seismic signals and is less susceptible to noise interference. For example, when processing seismic signals with substantial noise, the KAN module can leverage its unique nonlinear fitting ability to extract key seismic features from the noisy signals and accurately identify the patterns of seismic signals.

3.3. Hybrid position encoding

In seismic signals, both frequency-domain and time-domain features play significant roles in classification tasks. Therefore, we propose a hybrid position encoding method to enhance the model's representation capability in both frequency and time domains. This position encoding method combines spectral encoding and temporal encoding, providing the model with richer signal representations.

3.3.1. Spectral position encoding

Spectral position encoding enhances the model's sensitivity to the spectral features of seismic signals by encoding frequency positions. The calculation formula for spectral encoding is as follows:

$$PE_{freq}(pos,\ 2i) = \sin \frac{pos}{10000^{2i/d_{model}}} \tag{6}$$

This formula encodes frequency positions through sine functions, aiding the model in focusing on variations of different frequency components in seismic signals. Spectral position encoding provides the model with spatial information in the signal's frequency domain, facilitating the extraction of frequency features. For example, seismic waves of different frequencies may correspond to distinct geological structures or types of seismic activities. Spectral position encoding helps the model accurately distinguish these frequency features, thereby enhancing its understanding of the frequency- domain characteristics of seismic signals.

DOI: 10.54254/2755-2721/2025.CH24699

3.3.2. Temporal position encoding

Temporal position encoding is designed to capture the time-varying characteristics of seismic signals, particularly the propagation times of P-waves and S-waves. The calculation formula for temporal position encoding is:

$$PE_{time}(t) = \log(1 + \tau \bullet t) \tag{7}$$

Where τ denotes a learnable scaling factor and t represents the time step. This encoding method encodes time through logarithmic mapping, enabling adaptation to the temporal evolution characteristics of seismic signals. Given that the temporal patterns of seismic signals often exhibit nonlinearity, logarithmic encoding can better represent such variations. For instance, during the propagation of seismic waves, the intervals between arrival times are non-uniformly distributed, and logarithmic encoding can more accurately reflect the nonlinear changes in these intervals, assisting the model in better capturing the temporal features of seismic signals.

3.4. Model architecture

The proposed KAN-Former architecture is built upon the Transformer framework with innovative enhancements. The processing workflow of the entire model architecture is as follows:

1. Input Signal: Raw seismic signal data is input and, after hybrid position encoding, serves as the input to the Transformer model. The hybrid position encoding integrates frequency- domain and time-domain positional information into the signal, enabling the model to better understand features across different dimensions.

2. Hierarchical Attention Mechanism: The signal first undergoes global feature extraction through the global attention layer, followed by refinement of local time-domain features via the local attention window mechanism. The global attention layer captures the overall patterns of the signal, while the local attention window focuses on critical local variations. Together, they provide rich feature information for subsequent processing.

3. KAN-Former Module: Within each layer of the Transformer structure, the traditional feedforward network (FFN) is replaced by the KAN-Former module. This module captures complex patterns in seismic signals through nonlinear transformations, effectively enhancing the model's ability to process intricate seismic waveforms.

4. Output Layer: After the aforementioned processing steps, the model outputs classification results for the seismic signals via the classification layer. The classification layer leverages the feature information extracted and processed by preceding layers to accurately categorize seismic signals, determining attributes such as earthquake type and magnitude range.

The advantage of this architecture lies in its ability to balance both global and local features of seismic signals while leveraging the powerful expressive capabilities of the KAN network to better fit nonlinear features in seismic data. Ultimately, this model not only improves classification accuracy but also enhances the interpretability of seismic signal classification—particularly in waveform analysis and understanding wave propagation laws—providing a robust tool for in-depth seismic signal research.

DOI: 10.54254/2755-2721/2025.CH24699

4. Experiments

4.1. Experimental setup

The experiments utilize the STEAD dataset, which comprises 12,000 samples with a ratio of 7:3 between seismic signals and noise signals. To comprehensively evaluate the model's performance, the adopted metrics include Accuracy, F1 Score, Number of Parameters (Params), FLOPs (Floating Point Operations), and Training Time (minutes). All experiments were conducted on an NVIDIA A100 GPU.

4.2. Benchmark model comparison

Table 1. Performance comparison

Model	Acc(%)	F1	Params(M)	FLOPs(G)	Training Time (min)
1D-CNN	85.2	0.832	2.1	0.18	23
AlexNet-1D	82.7	0.798	3.9	0.24	27
VGG16-1D	88.4	0.862	48.6	3.21	65
LSTM(Bi)	86.9	0.852	8.3	1.02	41
Transformer	91.5	0.902	25.3	3.85	79
ResNet18	92.8	0.917	11.7	1.22	38
InceptionTim e	93.1	0.920	14.9	1.58	53
ResNet- KAN(original)	96.1	0.955	13.1	1.45	45
Ours	97.5	0.968	15.7	2.15	63

We first compared the classification performance of Transformer-KAN and the above baseline models on the test set. The results show that the Transformer-KAN model achieved an overall accuracy of 97.5%, which is better than 91.5% of the pure Transformer model and 90.4% of the CNN model. Especially for test samples with high noise proportion (signal-to-noise ratio less than 10 dB), Transformer-KAN can still maintain an accuracy of about 95%, while the pure Transformer and CNN drop to about 88% and 80% respectively. This indicates that the model in this paper has stronger robustness and discrimination under strong noise conditions. In addition, we calculated the precision, recall and F1 score of each model. Transformer-KAN obtained 98.1% precision and 97.4% recall in the earthquake event category, more than 95% in the noise category, and a comprehensive F1 score of 0.968, which is significantly higher than 0.931 of the pure Transformer and about 0.89 of the CNN.

This means that Transformer-KAN not only has higher overall accuracy, but also performs better in avoiding false positives and false negatives. The advantages of Transformer-KAN are also reflected in model efficiency and complexity. Since the KAN-Former module uses functional weights, its number of parameters is relatively sparse. According to statistics, the total number of parameters of Transformer-KAN is about 8.7 million, which is only about 60% of that of the pure Transformer and about 50% of that of the CNN. This simplification does not come at the cost of performance loss, but is accompanied by performance improvement. This confirms that KAN can effectively suppress parameter redundancy while improving the model's expression ability.

DOI: 10.54254/2755-2721/2025.CH24699

4.3. Ablation studies

Table 2. Performance changes

Configuration	Acc(%)	Δ Acc	F1
Base Transformer	91.5	-6.0↓	0.902
+ KAN-Former	95.2	-2.3↓	0.943
+ Dual Attention	96.3	-1.2↓	0.958
+ Seismic PE	96.9	-0.6↓	0.963
Full Model	97.5	—	0.968

To thoroughly evaluate the role of each model component, we conducted ablation experiments by sequentially removing or replacing key modules of Transformer-KAN and observing performance changes. First, we investigated the importance of the KAN-Former module. By replacing the KAN-Former in Transformer-KAN with an equivalent two-layer ReLU feedforward network (restoring it to a standard Transformer layer) while keeping other structures unchanged, we obtained Model Variant-K (without KAN). Test results showed that the overall accuracy of this variant dropped to 92.8%, a decrease of approximately 3.7 percentage points compared to the complete model. Among them, the recall for microseismic events decreased from 95% to 89%, indicating that the model without the KAN module was more prone to missing small seismic events, while the false alarm rate also increased. This phenomenon suggests that the KAN module plays a role in extracting key features and reducing noise interference. By flexibly adjusting edge functions, KAN may learn to suppress noise components and enhance signal components. When removed, the model's nonlinear fitting capability weakens, leading to performance degradation.

Next, we evaluated the contribution of the hierarchical attention mechanism. We modified the model to use single-layer global attention (directly applying the standard Transformer encoder to the complete sequence for several layers) and removed the local attention layer, resulting in Model Variant-H (without hierarchical structure). This variant achieved an accuracy of 91.5%, a decrease of more than 5 percentage points compared to the complete model. Analysis of error cases showed that Variant-H had significant difficulties in identifying seismic onset signals immediately following noise, often submerging short-term weak seismic phases in the overall waveform and thus missing events. This confirms the importance of hierarchical attention: without local attention to first extract detailed features within short time windows, global attention struggles to capture weak effective signals in a timely manner. In contrast, the complete model with a hierarchical structure can highlight event signs at the local level and then integrate information at the global level, thus achieving higher sensitivity and accuracy. Therefore, hierarchical attention significantly improves event detection rates in complex backgrounds.

Finally, we performed ablation tests on the hybrid position encoding, examining two scenarios: using only absolute position encoding and only relative position encoding. When using only absolute position encoding (Variant-A), the model accuracy dropped to 94.0%; when using only relative position encoding (Variant-R), the accuracy was 94.5%. Both showed slight decreases compared to the complete model, particularly demonstrating stability degradation in cross-record comparison and data with different sampling rates. Specifically, Variant-A had difficulty correctly focusing on key segments when encountering records with time-length distributions significantly different from those during training, presumably because the lack of relative position information made it difficult for the model to generalize across sequences of different lengths. Variant-R,

DOI: 10.54254/2755-2721/2025.CH24699

however, had deficiencies in locating the absolute occurrence time of events—for example, sometimes misjudging events in the beginning part of long records—indicating that the absence of global position references affected the model's understanding of the time origin. In contrast, the hybrid encoding strategy of the complete model performed well in both aspects: regardless of record length, the model could effectively use relative positions to capture local patterns, while global absolute positions provided temporal anchors, ensuring the model did not miss events that might exist at the beginning or end of sequences. Thus, the hybrid position encoding synergizes absolute and relative position information, achieving better results than single encoding in this task.

In summary, the ablation experiments clearly validate the necessity of each component module of the Transformer-KAN model. The KAN-Former module significantly contributes to classification performance by enhancing nonlinear fitting capability and parameter efficiency; the hierarchical attention effectively improves the model's event capture ability in noisy conditions; and the hybrid position encoding ensures the model's comprehensive perception of temporal order and relative relationships. These components collectively contribute to the high accuracy and robustness of Transformer-KAN.

4.4. Significance verification

Table 3. Friedman test（a=0.05）

Pair	p-value	Δ Acc
Ours vs ResNet-KAN	2.8e-5	+1.4%
Ours vs Inception	1.2e-6	+4.4%

A Friedman test (α=0.05) was conducted to verify the significance among all models, with results indicating statistically significant differences ($p=5.6\times10^{-11}$). Comparative analyses between the proposed method and ResNet-KAN, as well as InceptionTime, yielded p-values of 2.8×10^{-5} (accuracy improvement of 1.4%) and 1.2×10^{-6} (accuracy improvement of 4.4%), respectively. Furthermore, in a -5 dB Gaussian noise environment, the F1 score of this method remained at 92.3%, significantly higher than other models (all below 89%), demonstrating strong robustness in noisy conditions. The method achieved 90% accuracy within 22 training epochs, whereas the Transformer required 38 epochs, indicating faster convergence speed.

These results collectively confirm that Transformer-KAN not only achieves optimal performance in accuracy and F1 score but also demonstrates significant advantages in parameter efficiency, computational efficiency, and training efficiency, providing an effective and reliable solution for earthquake early warning systems.

5. Discussion

The Transformer-KAN architecture has made remarkable progress in addressing the challenges of seismic signal classification. Although the model's inference speed is 21% lower than that of ResNet- KAN, its capability to support parallel processing across multiple stations significantly enhances overall computational efficiency. This design is particularly crucial for real-time earthquake early warning systems, as it can effectively handle the processing requirements of multi-source data while ensuring accuracy. The weight distribution of the KAN layer exhibits a significant correlation with the attenuation law of seismic waves. This finding not only enhances the physical interpretability of the model but also provides a new analytical perspective for the propagation

DOI: 10.54254/2755-2721/2025.CH24699

mechanism of seismic waves, contributing to a deeper understanding of the physical characteristics of seismic signals.

Through model compression techniques, Transformer-KAN can achieve real-time processing on edge computing devices such as the Jetson Nano. This characteristic makes it highly suitable for deployment in resource-constrained seismic monitoring stations, offering a feasible solution for real-time earthquake early warning.

6. Conclusion

This paper proposes a Transformer-KAN model for robust seismic signal classification, which combines the Transformer's attention mechanism with the nonlinear representation capability of the Kolmogorov-Arnold network to form an innovative architecture integrating hierarchical attention extraction, multi-scale position encoding, and functional feedforward module fusion. Experiments on real seismic datasets demonstrate the model's significant advantages over traditional methods in seismic event detection and noise identification tasks. Transformer-KAN maintains high accuracy in strong noise environments, significantly reducing false alarm and miss rates, verifying the effectiveness of hierarchical attention and KAN modules in enhancing model discriminability.

Ablation studies further show that each key component (KAN-Former, hierarchical attention, hybrid position encoding) makes an indispensable contribution to model performance.

The main contributions of this study are: (1) proposing the Transformer-KAN framework, introducing learnable functional weights into the attention network to achieve more refined feature modeling of seismic waveforms; (2) designing hierarchical attention and hybrid position encoding to optimize the Transformer structure for the temporal scale diversity and positional dependency characteristics of seismic signals; (3) verifying through systematic experiments the method's robustness in complex noise backgrounds and classification performance superior to existing models. Given the pattern similarities between seismic signals and other time-series signals, the Transformer- KAN concept is also expected to be generalized to other fields requiring high-precision pattern recognition, such as structural health monitoring and acoustic event detection.

In future work, we will attempt to apply the model to larger-scale seismic monitoring network data and explore its deployment feasibility in real-time early warning systems. Additionally, leveraging the interpretability of KAN modules, we plan to deeply analyze the functions learned by the model to explore their physical meanings, aiming to provide new clues for seismological mechanism research. In summary, the Transformer-KAN model proposed in this paper provides a new tool for intelligent analysis of seismic signals, holding potential important application value in improving early warning reliability and disaster mitigation.

References

[1] LIU Xi-wu, LIU Hong, LI You-ming, NIAN Jing-bo. Local waves decomposition and its applications to seismic signal time-frequency analysis [J]. Progress in Geophysics, 2007, 22(2): 365-375.

[2] Huang H M, Bian Y J, Lu S J, et al. A wavelet feature research on seismic waveforms of earthquakes and explosions [J]. Acta Seismologica Sinica, 2010, 32(3): 270-276.

[3] Liu X W, Zhang N, Gou Y F, et al. The comparison and application of time-frequency analysis methods to seismic signal [J]. Progress in Geophysics (in Chinese), 2008, 23(3): 743-753.

[4] Zhao M, CHEN S, Dave Y. Waveform classification and seismic recognition by convolution neural network [J]. Chinese Journal of Geophysics, 2019, 62(1): 374-382.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

DOI: 10.54254/2755-2721/2025.CH24699

[6] Gao Z, Shi X, Wang H, et al. Earthformer: Exploring space-time transformers for earth system forecasting [J]. Advances in Neural Information Processing Systems, 2022, 35: 25390-25403.
[7] Liu Z, Wang Y, Vaidya S, et al. Kan: Kolmogorov-arnold networks [J]. arXiv preprint arXiv: 2404.19756, 2024.
[8] Mousavi S M, Ellsworth W L, Zhu W, et al. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking [J]. Nature communications, 2020, 11(1): 3952.