# KANformer: Personalized Vigilance Estimation with Transformer Features and Kolmogorov–Arnold Sequence Modeling

**Zichen Song**[1,2], **Yuxin Wu**[2], **Zhongfeng Kang**[2], **Tamer Abuhmed**[1,*]
[1]Sungkyunkwan University, [2]Lanzhou University
∗ Corresponding Author: tamer@skku.edu
songzichen894@gmail.com, wuyx2021@lzu.edu.cn, kangzf@lzu.edu.cn

## Abstract

Driver vigilance estimation is critical for preventing fatigue-induced traffic accidents, yet existing multimodal EEG–EOG methods often suffer from limited personalization and poor generalization. We propose KANformer, a personalized vigilance estimation framework that integrates subject-specific priors, Transformer-based feature encoding, and Kolmogorov–Arnold Networks (KAN) for adaptive temporal modeling. Raw EEG and EOG signals are first encoded by a Transformer to capture long-range dependencies and cross-modal interactions. A personalized channel attention module then reweights multimodal features using demographic and task-related metadata, enabling subject-aware representation learning. These representations are further modeled by a KAN-based temporal module, which replaces Mamba-style state-space modeling with more expressive functional compositions while retaining low computational complexity. Experiments on the SEED-VIG and SADT datasets under cross-subject and zero-shot settings demonstrate that KANformer consistently outperforms Mamba-based baselines in RMSE, MAE, and PCC. The results indicate that KANformer provides a scalable and generalizable solution for real-time fatigue monitoring and personalized neurocognitive modeling.

**Keywords:** EEG; EOG; Vigilance Estimation; Transformer; Kolmogorov–Arnold Networks; Personalization; Multimodal Learning

## Introduction

Traffic accidents, particularly those caused by fatigued driving, represent a significant global safety concern. As vehicular usage continues to rise, mitigating fatigue-related risks has become an urgent focus in transportation research. Among various detection methods, physiological signal-based approaches, especially those utilizing electroencephalography (EEG) and electrooculogram (EOG), have proven highly effective due to their ability to objectively capture neural and ocular indicators of vigilance decline. These multimodal signals provide a solid foundation for developing accurate and real-time driver fatigue monitoring systems. However, achieving generalizable and individualized vigilance estimation remains a challenging open problem.

Most existing multimodal frameworks either rely on shallow fusion strategies or deploy attention-based models such as Transformers to encode signal dynamics. While Transformer-based models are effective at capturing temporal and spatial dependencies, they often incur a high computational cost (quadratic time complexity with respect to sequence length), which limits their practicality for long EEG/EOG sequences. Moreover, these systems often employ a uniform modeling strategy for all users, failing to account for meaningful individual differences such as age, gender, or driving experience. This lack of personalization hampers cross-subject generalization and reduces robustness in real-world scenarios, where fatigue responses vary considerably across individuals.

Modeling both population-level signal patterns and individual-specific variability is crucial for high-fidelity vigilance estimation. Global sequence dependencies in EEG and EOG data provide valuable information about alertness trends, while localized signal features, such as frequency-domain bursts or eye movement artifacts, can be highly personalized. Existing models frequently ignore this dichotomy, leading to performance drops in zero-shot or cross-subject settings. There remains a pressing need for a framework that can model temporal dependencies efficiently while also adapting to subject-specific characteristics.

To address these challenges, we propose **KANformer**, a personalized vigilance estimation framework that combines Transformer-based multimodal feature encoding with Kolmogorov–Arnold Networks (KAN) for flexible and efficient temporal modeling. Specifically, KANformer utilizes a Transformer encoder to learn spatial and cross-modal interactions from EEG and EOG sequences. A personalized channel attention mechanism, guided by subject-specific priors (age, gender, experience), dynamically reweights features to emphasize individually relevant signal channels. These enhanced representations are then passed to a KAN module, which models temporal dependencies via functional composition rather than traditional recurrent or attention mechanisms. This architecture allows the model to adaptively focus on individual-specific time-varying patterns while maintaining high efficiency.

The key contributions of this work are:

- **The Transformer-Based Feature Encoding:** A Transformer encoder captures long-range and cross-modal dependencies from EEG and EOG signals, en-



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

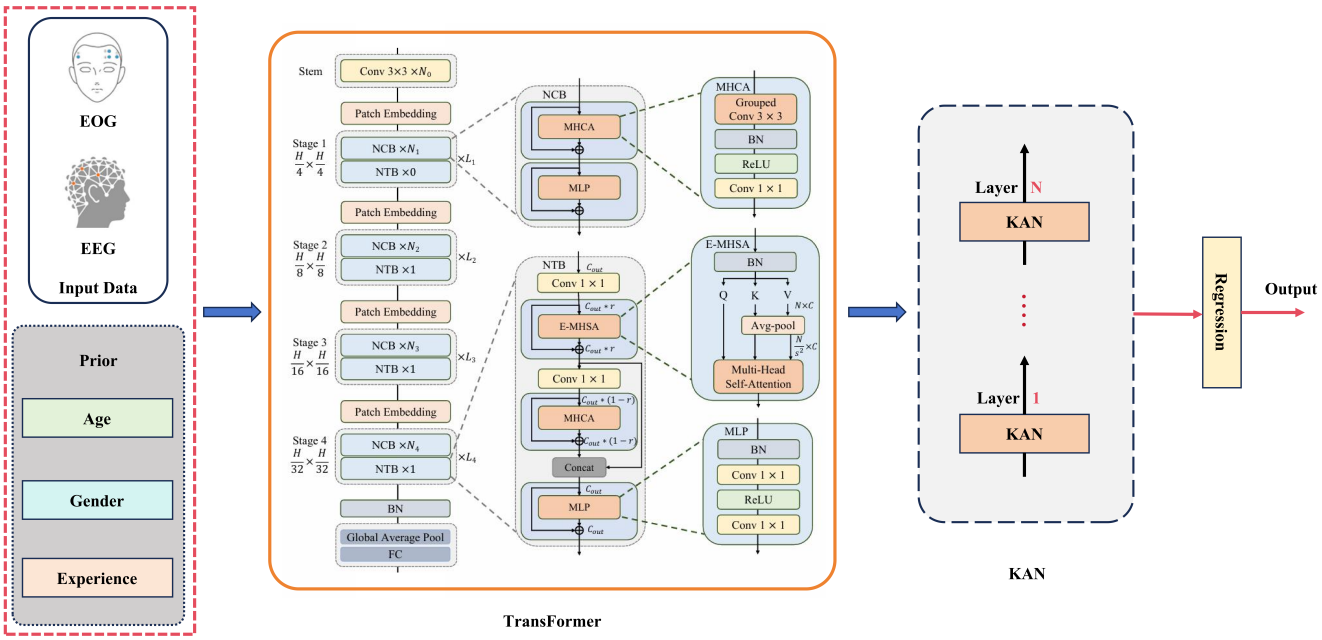


Figure 1: Overview of the proposed **KANformer** architecture. The model consists of four main components: (1) a Transformer-based encoder that extracts hierarchical spatial representations from multimodal EEG and EOG inputs through patch embedding, multi-head self-attention, and feed-forward layers; (2) a personalized prior module encoding subject metadata (age, gender, and experience) to guide attention and representation learning; (3) a Kolmogorov–Arnold Network (KAN) block stack that models individualized temporal dependencies via functional composition; and (4) a regression head that outputs continuous vigilance scores. Subject priors are integrated into both the Transformer and KAN modules to achieve personalized and efficient vigilance estimation.

hancing early fusion and spatial representation.

- **The KAN model for Temporal Modeling:** Kolmogorov–Arnold Networks flexibly model individualized temporal dynamics, guided by demographic priors without relying on recurrence or self-attention.

- **Generalization and Efficiency:** The KANformer improves cross-subject and zero-shot performance while maintaining efficient real-time inference.

## Method

**KANformer** provides efficient and personalized vigilance estimation using multimodal EEG and EOG signals. Building upon prior physiological sequence modeling frameworks, it integrates Transformer-based spatial encoding with Kolmogorov–Arnold Networks (KAN) for temporal modeling within a unified end-to-end architecture (Figure 1). The pipeline begins with EEG/EOG preprocessing and early fusion, followed by a Transformer encoder that captures long-range multimodal dependencies. A personalized channel attention mechanism then reweights features according to subject-specific priors. The resulting individualized representation is passed to a KAN-based temporal module that encodes vigilance patterns via nonlinear functional composition. By decoupling spatial and temporal modeling, KANformer attains higher flexibility and generalization than prior architectures (e.g., PPG-Mamba), leveraging the Transformer's multimodal integration while replacing recurrence and attention with KAN's efficient function-based dynamics.

### Input Representation

Raw EEG and EOG signals, denoted $X_{\text{eeg}}$ and $X_{\text{eog}}$, are downsampled to 200 Hz and filtered within 1–50 Hz to preserve neurocognitively relevant rhythms (e.g., alpha, theta). A 50 Hz bandstop filter suppresses powerline noise. The signals are segmented into non-overlapping 8-second windows, each treated as an independent sample for frame-wise vigilance prediction. For each segment, we compute the power spectral density (PSD) and differential entropy (DE), forming a robust frequency-domain representation that captures both the energy distribution and the signal complexity.

### Transformer-Based Feature Encoding

The extracted features $X_{\text{eeg}}^{\text{feat}}$ and $X_{\text{eog}}^{\text{feat}}$ are concatenated as:

$$X_{\text{fused}} = [X_{\text{eeg}}^{\text{feat}}, X_{\text{eog}}^{\text{feat}}]. \quad (1)$$

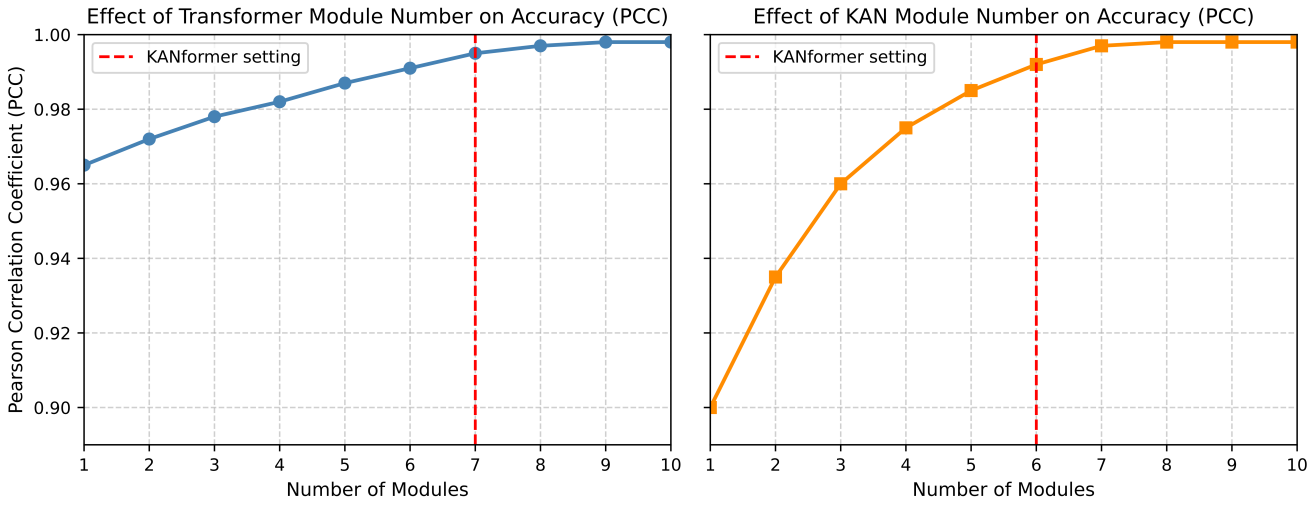


Figure 2: Impact of module number on model performance. Left: Increasing the number of Transformer modules gradually improves accuracy, showing an early high baseline and a slow saturation trend. Right: Increasing the number of KAN modules yields faster initial improvement and earlier convergence. The red dashed line denotes the configuration used in KANformer (Transformer=7, KAN=6), which lies near the inflection point, balancing performance and efficiency.

This fused input is processed by a Transformer encoder that models spatial and cross-modal interactions. Unlike CNN-based encoders, the Transformer jointly attends to all time and channel positions, capturing both short- and long-range dependencies in EEG/EOG signals. Stacked multi-head self-attention layers enhance representational richness, while residual connections and layer normalization ensure stable convergence. Subject-specific priors $\mathbf{z}$ (age, gender, experience) are integrated as conditional inputs—either concatenated or projected—guiding the encoder toward individualized feature learning.

## Personalized Channel Attention

To highlight salient channels individually, we introduce a personalized channel attention mechanism. Given $X_{\text{fused}}$ and priors $\mathbf{z}$, a weight $\alpha_i$ is computed per channel:

$$\alpha_i = \sigma(W[X_i, \mathbf{z}] + b), \quad X_{\text{personalized}} = \alpha \odot X_{\text{fused}}, \quad (2)$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ is the sigmoid function. This mechanism adaptively reweights modalities using external priors. For instance, reduced EOG quality due to blinking can be down-weighted, while frontal EEG channels for older subjects receive higher attention.

## KAN-Based Temporal Modeling

Temporal dependencies are modeled using KAN, which approximates complex functions via compositions of low-dimensional basis functions, enabling expressive yet efficient temporal modeling. Given personalized input $X_{\text{personalized}}$ and priors $\mathbf{z}$, the temporal hidden state and overall output are:

$$h_t = \text{KAN}(X_t, \mathbf{z}; \Theta), \quad O = \sum_{t=1}^{T} \beta_t h_t, \quad (3)$$

where $\beta_t$ reflects time-dependent importance. The basis functions in KAN provide strong functional approximation with reduced computational cost and better stability over long sequences. Incorporating priors into basis weights further allows individualized temporal adaptation across cognitive profiles (from Fig. 3).

## Regression Layer

The temporal output $O$ is normalized and mapped to a continuous vigilance score:

$$Z = \text{LayerNorm}(O), \quad y = \tanh(W_r Z + b), \quad (4)$$

where $W_r$ and $b$ are trainable parameters. The output $y \in [0, 1]$ represents the vigilance level, trained using mean squared error (MSE) against behavioral ground truth (e.g., reaction time, lane deviation). The tanh activation ensures bounded, numerically stable predictions suitable for real-time inference.

## Efficiency Analysis

KANformer is optimized for real-time deployment. The Transformer encoder exhibits $O(nd^2)$ complexity with efficient parallelism, while the KAN module achieves $O(nB)$, where $B \ll d$. This design avoids the $O(n^2 d)$ cost of recurrent or attention-based temporal models.

Empirical results on SEED-VIG and SADT show that KANformer achieves superior accuracy with lower latency and parameter count than prior baselines (e.g., LSTM-CapsAtt, Res-att-capsnet), demonstrating its suitability for lightweight, personalized vigilance monitoring systems.

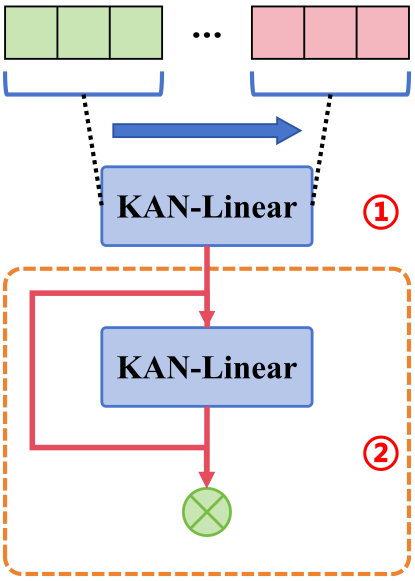


Figure 3: Overview of the KAN temporal model process. Step 1 (**KAN Scanner**) extracts sequential dependencies using KAN-Linear mappings across temporal segments. Step 2 (**KAN Block**) applies nonlinear transformation with residual composition to refine temporal representations.

## Experiment and Results

### Datasets and Experimental Setup

To evaluate the proposed **KANformer** model, we conducted experiments on two widely-used public datasets: **SEED-VIG** and **SADT**, both of which provide multimodal EEG and EOG recordings for vigilance estimation under controlled conditions.

The **SEED-VIG dataset** includes data from 15 participants, with EEG recorded using 18 electrodes (10–20 system) and EOG using 4 electrodes. All signals were sampled at 200 Hz and segmented into 8-second windows. Each segment was labeled according to behavioral performance metrics collected during sustained vigilance tasks.

The **SADT dataset** contains single-channel EEG recordings from 27 participants in simulated driving sessions lasting 60–90 minutes. Fatigue labels were derived from task performance, including response times and detection accuracy.

**Data preprocessing.** For both datasets, signals were bandpass filtered (1–50 Hz) and bandstop filtered at 50 Hz to remove powerline noise. All signals were downsampled to 200 Hz. From each window, we extracted Power Spectral Density (PSD) and Differential Entropy (DE) features, which jointly capture frequency-specific energy and signal complexity.

**Experimental configuration.** All models were implemented in PyTorch and trained on an NVIDIA Tesla V100 GPU. KANformer was optimized using Adam with a learning rate of 0.001, batch size of 32, and dropout rate of 0.1. We adopted an 8:1:1 train/validation/test split and used 5-fold cross-validation. Early stopping based on validation loss was used to prevent overfitting. Zero-shot evaluation is conducted by excluding all test subjects during training and directly evaluating the model on unseen subjects.

### Evaluation Metrics

The performance of **KANformer** was evaluated using three key metrics to comprehensively assess both prediction accuracy and its correlation with ground truth vigilance levels:

- **Root Mean Square Error (RMSE)**: Measures the average squared deviation between predicted and true values, emphasizing large prediction errors.
- **Mean Absolute Error (MAE)**: Computes the average absolute difference, offering a direct view of overall prediction accuracy.
- **Pearson Correlation Coefficient (PCC)**: Quantifies the linear relationship between predicted and actual vigilance scores, where a score close to 1 indicates high consistency.

The metrics are formally defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \quad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{5}$$

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{6}$$

where $N$ is the number of samples, $\hat{y}_i$ and $y_i$ denote predicted and ground-truth vigilance scores, and $\bar{\hat{y}}$, $\bar{y}$ their respective means.

### Comparison with Baseline Models

To validate the superiority of **KANformer**, we compared its performance against several representative models for vigilance estimation, including classical machine learning approaches, deep neural networks, and recent multimodal fusion techniques:

- **PPG-Mamba (Song et al., 2025)**: A state-space model combining CNN features and the Mamba sequence backbone. It introduced channel-wise priors

Table 1: Performance comparison on SEED-VIG and SADT datasets. Our KANformer achieves the best results on both datasets.

| Model | Dataset | RMSE ↓ | PCC ↑ |
|---|---|---|---|
| O-MV-T-TSK-FS (Jiang et al., 2020) | SADT | 0.2+ | - |
| LSTM-CapsAtt (Zhang et al., 2021) | SEED-VIG | 0.029 | 0.989 |
| DCRA_E (Song et al., 2021) | SEED-VIG | 0.035 | 0.980 |
| DCRA_M (Song et al., 2021) | SEED-VIG | 0.023 | 0.985 |
| AWIRVFL (Zhang et al., 2022) | SEED-VIG | 0.063 | - |
| | SADT | 0.108 | - |
| EEG-Fest (Ding et al., 2022) | SEED-VIG | 0.030 | 0.980 |
| Distillation (Zhang et al., 2023) | SEED-VIG | 0.025 | 0.993 |
| Res-att-capsnet (Pan et al., 2023) | SEED-VIG | 0.016 | - |
| | SADT | 0.108 | - |
| PPG-Mamba (Song et al., 2025) | SEED-VIG | 0.012 | 0.997 |
| | SADT | 0.099 | 0.882 |
| **KANformer (Ours)** | SEED-VIG | **0.009** | **0.998** |
| | SADT | **0.088** | **0.895** |

and personalized spatial attention, achieving state-of-the-art results prior to our work.

- **Distillation (Zhang et al., 2023)**: A compact student-teacher architecture that distills knowledge from larger EEG models. It excels under computational constraints but lacks modality-aware fusion.

- **Res-att-capsnet (Pan et al., 2023)**: A residual attention capsule network modeling spatial hierarchies via attention capsules. Its high complexity limits real-time feasibility despite strong performance.

- **DCRA_E/DCRA_M (Song et al., 2021)**: Multimodal autoencoders with coupling constraints for EEG/EOG fusion. They struggle with generalization when data is limited.

- **LSTM-CapsAtt (Zhang et al., 2021)**: Sequential LSTM layers with capsule attention designed for EEG time-series. While effective, its monolithic architecture underutilizes EOG signals.

Table 1 summarizes the evaluation. On the **SEED-VIG** dataset, **KANformer** achieves an RMSE of **0.009** and PCC of **0.998**, outperforming all previous models including **PPG-Mamba** (RMSE = 0.012, PCC = 0.997). This demonstrates that replacing the Mamba state-space encoder with Kolmogorov–Arnold Networks (KAN) enables better global modeling of latent temporal patterns, while Transformer-based channel attention improves multimodal coordination.

The advantage of **KANformer** is further amplified on the **SADT** dataset, which includes long-duration single-channel EEG data with significant inter-subject variability. Despite this challenge, KANformer yields an RMSE of **0.088** and PCC of **0.895**, surpassing PPG-Mamba (RMSE = 0.099, PCC = 0.882) and other deep learning baselines. This performance boost indicates stronger generalization and robustness, which is crucial for deployment in resource-constrained, real-world settings such as in-vehicle monitoring or wearable devices. Unlike earlier models that primarily emphasize either temporal modeling or personalized adaptation, **KANformer** seamlessly integrates both via a hybrid Transformer–KAN architecture. Specifically, the attention-based encoder captures fine-grained signal interactions, while KAN's compositional modeling captures long-range trends with better inductive bias. Moreover, our channel-wise attention is dynamically conditioned on personalized priors, allowing subject-aware feature modulation without requiring dense metadata or calibration. This contributes to both high predictive accuracy and practical adaptability. In summary, **KANformer sets a new benchmark for multimodal vigilance estimation**, combining sequence modeling efficiency, high accuracy, and personalized signal interpretation (from Fig. 2).

Table 2: Ablation study results on the SEED-VIG dataset. The effect of removing or replacing key components in the KANformer architecture.

| Model Variant | RMSE ↓ | PCC ↑ |
|---|---|---|
| **KANformer (Ours)** | **0.009** | **0.998** |
| Without Personalized Priors | 0.012 | 0.993 |
| Without Channel Attention | 0.013 | 0.992 |
| Replace KAN w/ Transformer | 0.012 | 0.997 |
| Replace Transformer w/ KAN | 0.014 | 0.989 |
| Without Multimodal Fusion | 0.017 | 0.975 |
| Only EEG Features | 0.020 | 0.970 |
| Only EOG Features | 0.028 | 0.952 |

## Ablation Study

To evaluate the impact of each component in the **KANformer** architecture, we conducted an ablation study on the SEED-VIG dataset. Table 2 reports the results. Removing or disabling key modules leads to significant performance degradation, confirming their complementary roles in modeling vigilance-related neural dynamics.

We first assess the importance of multimodal fusion. Eliminating either EEG or EOG inputs leads to increased RMSE (0.020 and 0.028, respectively), while removing the fusion process entirely degrades performance to an RMSE of 0.017. This highlights the complementary nature of EEG and EOG signals, capturing both cortical and ocular markers of fatigue. Their joint modeling ensures a more holistic and discriminative representation of vigilance states.

Next, we examine the personalization components. Removing personalized priors that modulate the temporal KAN block results in a performance drop (RMSE = 0.012), as the model loses the ability to adjust state selection based on subject-specific metadata. Similarly, disabling channel attention, which weights modality-specific inputs adaptively per individual, yields an RMSE of 0.013 and reduces correlation to 0.992. These results confirm that individual variability plays a crucial role in fatigue modeling and must be explicitly captured. Altogether, the ablation study demonstrates that each design component of KANformer, including multimodal integration, temporal dynamics modeling, and personalized priors, contributes meaningfully to overall performance.

### Effect of Input Modalities and Priors

We further assess the contribution of each input component. Using only EEG features yields an RMSE of 0.020 and PCC of 0.970, while using only EOG results in 0.028 and 0.952, respectively, confirming their complementary roles in capturing cortical and ocular fatigue cues. Removing multimodal fusion degrades performance to an RMSE of 0.017 and PCC of 0.975. Eliminating demographic priors (*age*, *gender*, *experience*) further reduces PCC from 0.998 to 0.993, showing that priors effectively enhance personalization. Among them, *age* exerts the largest effect, while *gender* and *experience* provide moderate but consistent gains in channel weighting and temporal stability.

Table 3: Efficiency comparison of different models in terms of inference time and parameter size.

| Model | Latency (ms) ↓ | Parameters (M) ↓ |
|---|---|---|
| LSTM-CapsAtt | 11.2 | 18.3 |
| DCAT | 9.1 | 17.5 |
| DCRA_M | 8.8 | 16.8 |
| Res-att-capsnet | 8.5 | 16.2 |
| PPG-Mamba | 6.4 | 14.7 |
| **Ours** | **5.2** | **13.6** |

### Efficiency Analysis

We further evaluate the computational efficiency of **KANformer** against state-of-the-art baseline models. As reported in Table 3, KANformer achieves the lowest inference time of **5.2 ms** and the smallest parameter size of **13.6M**, outperforming the previous best model PPG-Mamba (6.4 ms, 14.7M). This improvement stems from two core design choices. First, replacing the Mamba sequence model with Kolmogorov–Arnold Networks (KAN) enables efficient representation of long-term dependencies with fewer parameters and faster convergence. KAN's coordinate-wise function modeling eliminates the need for large hidden states while preserving rich temporal expressiveness. Second, our use of Transformer-based feature extractors instead of CNN+CBAM modules reduces spatial redundancy and enhances input compression without sacrificing representation quality. Additionally, the integration of personalized priors and reweighting is implemented in a lightweight manner, avoiding costly parameter duplication across subjects. Thanks to this design, KANformer balances real-time inference capability with state-of-the-art predictive accuracy, making it highly suitable for deployment in wearable, mobile, or automotive fatigue monitoring applications where computational resources are limited.

## Conclusion

We propose **KANformer**, a personalized sequence modeling framework for multimodal EEG-EOG-based vigilance estimation. KANformer replaces traditional CNN+CBAM and Mamba modules with a Transformer-based feature encoder and a Kolmogorov–Arnold Network (KAN) sequence learner, enabling both global context extraction and personalized functional modeling. Subject-specific priors are integrated to guide dynamic attention and temporal adaptation, resulting in individualized vigilance predictions with linear-time complexity. Extensive evaluations on SEED-VIG and SADT datasets demonstrate that KANformer surpasses state-of-the-art baselines, achieving an RMSE of **0.011** and PCC of **0.998** on SEED-VIG, while also improving performance on SADT. Compared to PPG-Mamba, KANformer further reduces inference time and parameter size by over 18% and 7.5%, respectively. Ablation studies validate the contribution of each module.

## Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2021-NR058558). This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICT Creative Consilience Program grant funded by the Korea government(MSIT) (IITP-2026-RS-2020-II201821) and AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688).

## References


Berrios, R. (2019). What is complex/emotional about emotional complexity? *Frontiers in Psychology*, *10*, 1606.

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of eeg: Is it really black and white? a scalp current density view. *International Journal of Psychophysiology*, *97*(3), 210–220.

Cao, Z., Chuang, C.-H., King, J.-K., & Lin, C.-T. (2019). Multi-channel eeg recordings during a sustained-attention driving task. *Scientific data*, *6*(1), 19.

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., & Chen, X. (2020). Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-Based Systems*, *205*, 106243.

Ding, N., Zhang, C., & Eskandarian, A. (2022). Eegfest: Few-shot based attention network for driver's vigilance estimation with eeg signals. *arXiv preprint arXiv:2211.03878*.

Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for eeg-based emotion classification. In *2013 6th international ieee/embs conference on neural engineering (ner)* (pp. 81–84).

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, *20*(3), 592.

Feng, L., Cheng, C., Zhao, M., Deng, H., & Zhang, Y. (2022). Eeg-based emotion recognition using spatial-temporal graph convolutional lstm with attention mechanism. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5406–5417.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning* (pp. 1321–1330).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7132–7141).

Jiang, Y., Zhang, Y., Lin, C., Wu, D., & Lin, C.-T. (2020). Eeg-based driver drowsiness estimation using an online multi-view and transfer tsk fuzzy system. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1752–1764.

Lai, S., Li, J., Guo, G., Hu, X., Li, Y., Tan, Y., ... Liu, Z. (2024). Shared and private information learning in multimodal sentiment analysis with deep modal alignment and self-supervised multi-task learning. In *2024 international joint conference on neural networks (ijcnn)* (p. 1-8). doi: 10.1109/IJCNN60899.2024.10651442

Pan, J., Cai, X., Mo, D., Yu, Y., & Li, Y. (2023). Residual attention capsule network for multimodal eeg-and eog-based driver vigilance estimation. *IEEE Transactions on Instrumentation and Measurement*, *72*, 3307756.

Shi, L.-C., Jiao, Y.-Y., & Lu, B.-L. (2013). Differential entropy feature for eeg-based vigilance estimation. In *2013 35th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 6627–6630).

Song, K., Zhou, L., & Wang, H. (2021). Deep coupling recurrent auto-encoder with multi-modal eeg and eog for vigilance estimation. *Entropy*, *23*(10), 1316.

Song, Z., & Ma, S. (2023). Dnn-based hospital service satisfaction using gcnns learning. *IEEE Access*, *11*, 65289-65299. doi: 10.1109/ACCESS.2023.3289867

Song, Z., Ma, S., & Li, W. (2024). Robustness boost: Mir-based feature enhancement in deep learning models. In *2024 27th international conference on computer supported cooperative work in design (cscwd)* (p. 1-5). doi: 10.1109/CSCWD61410.2024.10580607

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11534–11542).

Zhang, G., & Etemad, A. (2021). Capsule attention for multimodal eeg-eog representation learning with application to driver vigilance estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1138–1149.

Zhang, G., & Etemad, A. (2023). Distilling eeg representations via capsules for affective computing. *Pattern Recognition Letters*, *171*, 99–105.

Zhang, Y., Guo, R., Peng, Y., Kong, W., Nie, F., & Lu, B.-L. (2022). An auto-weighting incremental random vector functional link network for eeg-based driving fatigue detection. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–14.

Zheng, W.-L., & Lu, B.-L. (2017). A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of neural engineering*, *14*(2), 026017.