Efficient Modality Fusion Framework for Driver Cognitive Load Classification

Sumin Park *†, Sungjun Wang*, Chi Yoon Jeong*
*Digital Convergence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea
†Department of Biomechatronics Engineering, Sungkyunkwan University, Suwon, Republic of Korea
Email: taerang01@g.skku.edu, sungjoonkk@etri.re.kr, iamready@etri.re.kr

***Abstract*—Accurate estimation of drivers' cognitive load is essential for ensuring driving safety and performance. Although unimodal physiological signals, such as EEG or ECG, are widely used, they often provide limited information, leading to suboptimal classification performance. To address this limitation, we proposed an efficient modality fusion framework that leverages multimodal physiological signals for cognitive load classification. The framework can extract features from EEG and ECG signals using parallel 1*D* convolutional layers and ResNet-style blocks, followed by self-attention to refine intra-modality dependencies and cross-attention to capture complementary inter-modality interactions. Experiments on the CL-Drive public benchmark dataset evaluated the framework's performance under both binary and ternary classification settings, using 10-fold cross-validation and leave-one-subject-out (LOSO) protocols. The proposed framework consistently outperformed conventional machine learning models and state-of-the-art deep learning approaches, achieving accuracies of 85.69% (10-fold CV) and 76.26% (LOSO) for binary classification, and 78.79% (10-fold CV) and 63.68% (LOSO) for ternary classification. These results highlight the importance of attention-based multimodal fusion for robust cognitive load estimation, suggesting its strong potential for applications in intelligent transportation systems and brain–computer interface development.**

***Index Terms*—cognitive load classification, modality fusion, physiological signal analysis, mental state detection**

I. Introduction

Assessing the cognitive load provides valuable insights into users' mental states during complex tasks [1]. In particular, the cognitive load emerging from driving-related factors, such as navigation operations or drowsiness, has been strongly associated with traffic accidents [2], as it often causes delayed reactions and erroneous decision-making. Accordingly, understanding drivers' cognitive load is indispensable for ensuring both safety and driving performance [3]. Consequently, there is an increasing demand for research that aims to accurately classify drivers' cognitive load.

Methods for classifying driver cognitive load rely mainly on physiological signals, as these signals offer direct insights into driver mental states and facilitate precise classification [4]. Several deep- learning models employing unimodal physiological data have been proposed for analyzing driver states [5], [6]. Cui et al. [5] introduced an explainable convolutional neural network (CNN) model that integrates heterogeneous style features across subjects and minimizes interlabel distances, thereby mitigating intersubject variability in physiological signals. Pulver et al. [6] proposed a binary classification model for the cognitive load using unimodal electroencephalogram (EEG) data, leveraging a transformer architecture combined with transfer learning. Although unimodal approaches have been extensively conducted, such signals inherently provide limited information, hindering the ability to capture a comprehensive representation of the driver's state.

Consequently, recent studies have increasingly focused on integrating two or more multimodal physiological signals to achieve a holistic analysis of driver conditions and support various scenarios [7], [8], [9]. Liu et al. [7] combined EEG and Galvanic Skin Response (GSR) signals to improve emotion recognition. They first encoded each signal using multiple transformer encoders, and then fused the encoder outputs based on inter-modality interactions. In addition, they employed knowledge distillation to transform multimodal features into a unimodal GSR model, thereby enhancing the performance of the model for emotion classification. Azizi et al. [8] proposed an input-level fusion approach to assess the cognitive load of drivers by jointly analyzing EEG and ECG signals. They employed a pretrained biosignal transformer model [10] and fed it with integrated EEG and ECG signals. UniPhyNet was proposed by [9] to classify drivers' cognitive load using various physiological signals, such as EEG, ECG, and Electrodermal Activity (EDA). Each signal is encoded using a feature extraction block consisting of 1*D* convolutional blocks and a ResNet-style network, after which the features from each block are concatenated. Although multimodal deep learning models for physiological signal analysis offer advantages, such as greater robustness to noise, resilience to missing data, and a more comprehensive evaluation of drivers' cognitive states, they often face challenge: their classification performance is frequently inferior to that of unimodal approaches.

Thus, in this study, we address these limitations by proposing an efficient modality fusion method for driver cognitive load classification. Specifically, the method first encodes features from each physiological signal, and then employs self- and cross-attention mechanisms to model nonlocal temporal dependencies within each modality and facilitate the exchange of complementary information across modalities. This design mitigates the performance degradation often observed in multimodal settings as well as enhances the model's ability to capture complementary intermodal information, thereby

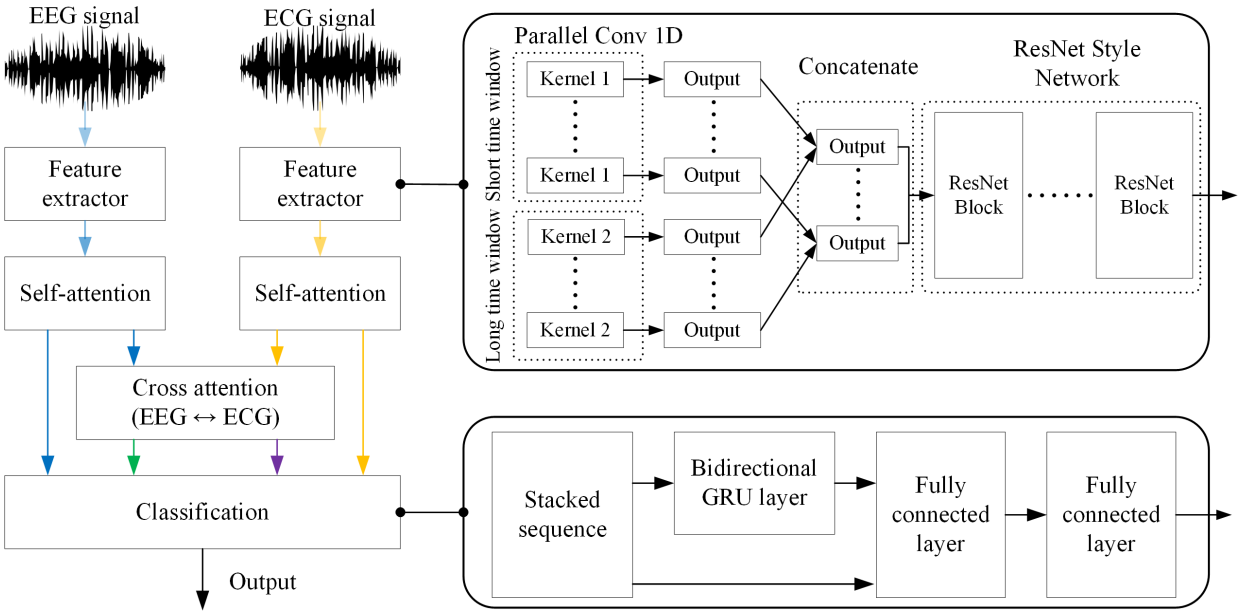

Fig. 1. Overview of the model architecture.

improving the overall classification accuracy and robustness.

The main contributions of this study are summarized as follows:

- We introduce an efficient modality fusion framework for driver cognitive load classification. The framework extracts features from multimodal physiological signals using a combination of 1*D* convolutional layers and ResNet-style networks, and captures complementary inter-modal dependencies through self- and cross-attention mechanisms.
- We demonstrate the effectiveness of the proposed framework on the CL-Drive public benchmark dataset [11], where it consistently outperforms unimodal baselines in terms of both accuracy and robustness.

The remainder of this paper is structured as follows: Section 2 introduces the proposed methodology, Section 3 details the experimental setup, Section 4 discusses the results, and Section 5 concludes the paper.

II. Proposed Method

We introduced an efficient modality fusion framework to classify drivers' cognitive load. The proposed method employs two modalities, EEG and ECG signals as inputs. The architecture of the model is illustrated in Fig. 1 comprises three main stages: the feature extractor, modality fusion module, and classification module.

A. Feature Extractor

The feature extractor, inspired by UniPhyNet [9], encodes useful information from each signal and consists of 1*D* parallel convolutional blocks and a ResNet-style network, as shown in Fig. 1. Each input branch begins with several 1*D* convolutional layers running in parallel, and each layer employs a different kernel size. This design allowed the network to capture both rapidly changing local signal patterns and slowly varying temporal trends. By concatenating the resulting feature maps, the model formed a rich representation that integrates information across multiple timescales. Sigmoid-weighted linear unit (SiLU) [12] activation was employed to ensure stable gradients and improve optimization.

Subsequently, the multiscale features are processed using a stack of ResNet blocks [13] combined with a convolutional block attention module [14], which further enhances feature extraction. Within each residual block, the attention mechanism adaptively emphasizes informative signal components. This mechanism operates across both feature channels and along the temporal dimension, thereby guiding the network toward salient patterns while reducing the impact of irrelevant noise.

B. Modality Fusion Module

In this work, to simultaneously incorporate both the temporal order and modality-specific rich features of EEG and ECG, each unimodal feature sequence was represented as a two-dimensional matrix across the temporal and feature dimensions. Accordingly, the EEG and ECG sequences are denoted as:

$$\mathbf{X}_{\text{EEG}} \in \mathbb{R}^{T \times d_e}, \quad \mathbf{X}_{\text{ECG}} \in \mathbb{R}^{T \times d_c}, \tag{1}$$

where T represents the temporal length and d_e and d_c denote the embedding dimensions of the EEG and ECG modalities, respectively.

In the modality fusion module, unimodal feature sequences are first enriched through self-attention, which captures long-range temporal dependencies and refines modality-specific representation. The multi-head attention mechanism serves as

a sub-layer in both the encoder and decoder blocks of the transformer architecture and builds upon scaled dot-product attention [15]. The scaled dot-product attention can be formulated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \quad (2)$$

The resulting single-head attention function is expressed as follows:

$$\text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right). \quad (3)$$

In this case, each head i applies its own learnable weight matrices to transform $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ allowing it to focus on learning from a restricted subspace of the feature representation [15].

The outputs of all attention heads are concatenated along the feature dimension and subsequently projected with the parameter matrix $\mathbf{W}_o$ to form the Multi-Head self-attention (MHA) representation, defined as

$$\text{MHA}_{\text{self}}(\mathbf{X}) = \text{Concat}\big(\text{Attn}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h)\big)_{h=1}^{H}\mathbf{W}_o. \quad (4)$$

This operation enables the model to learn nonlocal temporal dependencies that cannot be captured by only convolutional filters.

Considering $\mathbf{X}_{\text{EEG}}$ and $\mathbf{X}_{\text{ECG}}$, the refined unimodal representations $\tilde{\mathbf{X}}_{\text{EEG}}$ and $\tilde{\mathbf{X}}_{\text{ECG}}$ are computed by applying MHA as follows:

$$\tilde{\mathbf{X}}_{\text{EEG}} = \text{MHA}_{\text{self}}(\mathbf{X}_{\text{EEG}}), \quad \tilde{\mathbf{X}}_{\text{ECG}} = \text{MHA}_{\text{self}}(\mathbf{X}_{\text{ECG}}). \quad (5)$$

Subsequently, to enable the exchange of complementary information across modalities, bidirectional cross attention was applied, yielding the following formulation:

$$\hat{\mathbf{X}}_{\text{EEG}} = \text{MHA}_{\text{cross}}\left(\tilde{\mathbf{X}}_{\text{EEG}}\mathbf{W}_q^e, \tilde{\mathbf{X}}_{\text{ECG}}\mathbf{W}_k^c, \tilde{\mathbf{X}}_{\text{ECG}}\mathbf{W}_v^c\right), \quad (6)$$

$$\hat{\mathbf{X}}_{\text{ECG}} = \text{MHA}_{\text{cross}}\left(\tilde{\mathbf{X}}_{\text{ECG}}\mathbf{W}_q^c, \tilde{\mathbf{X}}_{\text{EEG}}\mathbf{W}_k^e, \tilde{\mathbf{X}}_{\text{EEG}}\mathbf{W}_v^e\right), \quad (7)$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ denote the learnable projection matrices for the query, key, and value, respectively, specific to each modality. This bidirectional interaction allows each modality to reinforce its representation by employing complementary information.

Finally, refined unimodal representations $\tilde{\mathbf{X}}_{\text{EEG}}$ and $\tilde{\mathbf{X}}_{\text{ECG}}$ obtained from self-attention, together with the cross-attention–enhanced features $\hat{\mathbf{X}}_{\text{EEG}}$ and $\hat{\mathbf{X}}_{\text{ECG}}$. are concatenated to form the final fused feature representation.

$$\mathbf{X}_{\text{fused}} = \text{Concat}\big[\tilde{\mathbf{X}}_{\text{EEG}}, \tilde{\mathbf{X}}_{\text{ECG}}, \hat{\mathbf{X}}_{\text{EEG}}, \hat{\mathbf{X}}_{\text{ECG}}\big]. \quad (8)$$

This fused representation integrates modality-specific refinements and cross-modal complementary information, thereby providing a comprehensive feature space for subsequent classification.

C. Classification Module

The fused representations are passed into a bidirectional GRU (Bi-GRU) to capture sequential dependencies over time explicitly:

$$\overrightarrow{h}_t = \text{GRU}(\mathbf{F}_t, \overrightarrow{h}_{t-1}), \quad \overleftarrow{h}_t = \text{GRU}(\mathbf{F}_t, \overleftarrow{h}_{t+1}), \quad (9)$$

$$h_t = [\overrightarrow{h}_t \| \overleftarrow{h}_t], \quad (10)$$

where $\mathbf{F}_t$ denotes the fused feature at the time step t. represents the current state. Accordingly, $\mathbf{h}_t$ is the feature representation at time step t, which is updated by integrating the current input $\mathbf{F}_t$ with the past and future hidden states [16].

This representation allows the Bi-GRU to aggregate contextual information from both past and future directions, enabling the network to capture long-term temporal dynamics. Consider the inherently time-varying nature of EEG and ECG signals, incorporating future context enhances the interpretation of the current state.

The output of the GRU is concatenated with the pooled attention features, providing a comprehensive multimodal representation. This concatenated vector is passed through a fully connected layer with an SiLU activation function [12] to introduce nonlinear transformations, thereby enhancing the feature expressiveness. Finally, a second fully connected layer projects the transformed representation onto the target label space to produce the classification logits.

III. Experimental Setup

The CL-drive dataset proposed by Angkan et al. [17] was used to evaluate the performance of the proposed method. This dataset was collected in a driving simulation environment to assess cognitive load, comprising physiological signals, such as EEG, ECG, and EDA, as well as eye-tracking data from 23 participants (17 females and 6 males). In our experiments, we utilized two physiological signals, EEG and ECG. The EEG data were recorded from four channels ('AF7', 'AF8', 'TP9', and 'TP10'), while the ECG data were obtained from three calibrated channels ('LL-RA', 'LA-RA', and 'Vx-RL'). After each driving session, the participants rated their perceived cognitive load on a 9-point Likert scale ranging from 1 (low) to 9 (high).

The raw ECG data contain numerous missing values, which hinder accurate model training. Therefore, we adopted the ECG preprocessing protocol proposed by Azizi et al. [8], in which a fifth-degree polynomial was fitted to corrupted channels, and the periodic pattern was reconstructed by replacing each missing value at time t with the corresponding value in the same phase.

All data samples were segmented into 10-second intervals, and for EEG, any segment containing missing values was discarded. During preprocessing, the sampling rates were set to 512 Hz and 256 Hz for the ECG and EEG signals, respectively.

For labeling, the self-reported Likert scores were mapped as follows: in the binary classification setting (following [8],

[9]), scores from 1–4 were labeled as 'low' and scores from 5–9 as 'high'; in the ternary classification setting, scores from 1–3 were labeled as 'low', scores from 4–6 as 'medium', and scores from 7–9 as 'high'. These binary and ternary labels served as the ground truth for model training and evaluation.

After preprocessing, the number of generated data samples was balanced to 3,074 for both the EEG and ECG modalities. Averagely, each participant contributed approximately 146 samples. The label distributions for binary and ternary classification are summarized in Table I.

TABLE I
LABEL DISTRIBUTION FOR BINARY AND TERNARY CLASSIFICATION TASKS

Setup	Label	Count
Binary classification	Low	1,302
	High	1,772
Ternary classification	Low	821
	Medium	1,604
	High	649

The EEG and ECG signals used in this study exhibit substantial inter-subject variability and are influenced by external factors, such as environmental conditions and sensor placement, which can alter signal phase [18], [19]. To consider these characteristics, we adopted both 10-fold cross-validation (10-fold) and Leave-One-Subject-Out (LOSO) evaluation protocols. A 10-fold setup was employed to evaluate the average performance of the model, whereas the LOSO protocol was used to assess the generalization capability across different subjects.

In all experiments, the proposed model was trained for 60 epochs with a batch size of 64 using the AdamW optimizer. Overfitting was observed early in training, as validation accuracy consistently lagged behind training accuracy. Therefore, we employed the ReduceLROnPlateau scheduler, which dynamically adjusts the learning rate based on improvements in validation performance. For data augmentation, three techniques—Gaussian noise, temporal warping, and random amplitude scaling—were applied following the approach of a previous study [9].

For EEG signal analysis in the feature extractor, kernel sizes of 5 and 11 were used in the $1D$ convolutional layers to capture the short- and long-time window features, respectively. For the ECG signal analysis, kernel sizes of 3 and 9 were employed. The outputs of these convolutional layers were encoded using nine and eight residual blocks for the EEG and ECG, respectively, thereby enabling deeper representation learning. In the modality fusion module, the embedding dimension was set to 64 and the number of heads was set to 4 for the self- and cross-attention mechanisms. For cognitive load classification, the GRU layer was configured with a hidden size of 128, followed by two fully connected layers of 128 neurons each.

IV. EXPERIMENTAL RESULTS

In this section, we present the evaluation results of the proposed model across four experimental settings: LOSO binary, LOSO ternary, 10-fold binary, and 10-fold ternary cross validations. To assess the model performance, we conducted comparisons against both classical machine learning and modern deep learning baselines. Classical baselines include Random Forest (RF) and Extreme Gradient Boosting (XGB), whereas deep learning baselines include ResNet, UniPhyNet [9], and transformer-based models [8]. Accuracy and F1-score were employed as evaluation metrics.

A. Effect of Modality Fusion

The results under the three configurations—unimodal EEG, unimodal ECG, and bimodal EEG–ECG—using both LOSO and 10-fold cross-validation are summarized in Tables II and III. Table II reports binary classification results, whereas Table III presents ternary classification results. In Table II, the bimodal EEG–ECG setting achieves the best performance under both 10-fold CV and LOSO, with accuracies of 85.69% and 76.26%, respectively. Similarly, in Table III, the bimodal configuration substantially outperforms unimodal EEG and ECG, achieving accuracies of 78.79% and 63.68% under the 10-fold CV and LOSO, respectively.

These results demonstrate that the integration of self-attention and cross-attention mechanisms enables a more effective fusion of modality-specific features. By leveraging complementary information across distinct physiological modalities, the proposed model achieved an accurate and a robust estimation of drivers' cognitive states.

TABLE II
PERFORMANCE OF CL-DRIVE BINARY CLASSIFICATION IN 10-FOLD CV AND LOSO TEST SETUPS

Setup	Modality	Accuracy (%)	F1-score (%)
10-fold CV	EEG	76.55	76.62
	ECG	84.52	84.57
	EEG & ECG	85.69	85.70
LOSO	EEG	62.22	61.48
	ECG	56.73	57.01
	EEG & ECG	76.26	76.55

TABLE III
PERFORMANCE OF CL-DRIVE TERNARY CLASSIFICATION IN 10-FOLD CV AND LOSO TEST SETUPS

Setup	Modality	Accuracy (%)	F1-score (%)
10-fold CV	EEG	67.93	67.96
	ECG	77.91	77.96
	EEG & ECG	78.79	78.77
LOSO	EEG	45.93	47.58
	ECG	43.24	43.32
	EEG & ECG	63.68	63.86

B. Comparison with State-of-the-art Models

Furthermore, we compared the performance of the proposed method with those of state-of-the-art approaches [8], [9] and conventional baselines. The results presented in Tables IV and V demonstrate that our method consistently outperformed all competing models. For instance, in binary classification, it

achieved accuracies of 85.69% under 10-fold CV and 76.26% under LOSO, surpassing conventional machine learning models (RF, XGB, and ResNet) and advanced deep learning methods (UniPhyNet and Transformer-based models).

Although various deep learning models operating directly on raw signals without handcrafted features perform on par with or even below feature-engineered machine learning models, our approach leverages attention-based multimodal fusion to emphasize the most discriminative representations. This design consistently provided superior accurate results and F1-scores across the evaluation settings.

Importantly, the pronounced performance gains under the LOSO protocol highlight the robustness of our framework to substantial intersubject variability, underscoring its strong generalization capability. These findings confirm that attention-driven multimodal fusion is crucial for exploiting cross-modal complementarities and achieving reliable driver cognitive-load classification.

TABLE IV
COMPARISON OF MODEL CLASSIFICATION PERFORMANCE BASED ON EEG AND ECG MODALITIES FOR CL-DRIVE BINARY CLASSIFICATION TASKS

Setup	Models	Accuracy (%)	F1-score (%)
10-fold CV	RF [9]	79.34	76.27
	XGB [9]	82.95	81.25
	ResNet [9]	64.49	62.14
	UniPhyNet [9]	79.33	79.24
	Transformer-based [8]	83.54	85.96
	Ours	**85.69**	**85.70**
LOSO	RF [9]	65.76	56.84
	XGB [9]	66.61	60.53
	ResNet [9]	63.99	56.73
	UniPhyNet [9]	73.61	74.06
	Transformer-based [8]	65.15	66.74
	Ours	**76.26**	**76.55**

TABLE V
COMPARISON OF MODEL CLASSIFICATION PERFORMANCE BASED ON EEG AND ECG MODALITIES FOR CL-DRIVE TERNARY CLASSIFICATION TASKS

Setup	Models	Accuracy (%)	F1-score (%)
10-fold CV	RF [9]	68.41	68.42
	XGB [9]	70.78	71.01
	ResNet [9]	56.56	50.09
	UniPhyNet [9]	73.60	73.78
	Transformer-based [8]	75.57	75.48
	Ours	**78.79**	**78.77**
LOSO	RF [9]	40.15	37.54
	XGB [9]	40.06	38.11
	ResNet [9]	60.36	47.58
	UniPhyNet [9]	62.64	62.02
	Transformer-based [8]	61.81	61.79
	Ours	**63.68**	**63.86**

C. Ablation Study

An ablation study was conducted to investigate the contribution of each component to the proposed fusion framework. Specifically, we examined four configurations: (i) a baseline feature-level fusion model without attention, (ii) a variant incorporating only self-attention, (iii) a variant incorporating only cross attention, and (iv) a full model that integrates both cross- and self-attention. This analysis allowed us to determine the relative importance of attentional mechanisms in enhancing multimodal feature integration.

The performances of these models were evaluated in terms of accuracy under a 10-fold CV setting, with results presented in Table VI. The model incorporating both self-attention and cross-attention achieved the best performance in both binary and ternary classification tasks. This improvement is attributed to the model's ability to refine intra-modality structures and enhance feature representations through self-attention, whereas cross-attention enables effective information exchange between modalities and reinforces complementary patterns across physiological signals.

TABLE VI
PERFORMANCE COMPARISON OF EEG AND ECG MODALITY FUSION STRATEGIES IN A 10-FOLD CROSS-VALIDATION SETUP

Setup	Fusion Strategy	Accuracy (%)
Binary classification	Simple feature-level	82.75
	Self-Attention–based	81.66
	Cross-Attention–based	83.50
	Self- and Cross-Attention–based	85.69
Ternary classification	Simple feature-level	77.78
	Self-Attention–based	77.92
	Cross-Attention–based	78.29
	Self- and Cross-Attention–based	78.79

V. CONCLUSION

In this study, we present an efficient modality fusion framework for driver cognitive load classification that integrates multiple physiological modalities to capture a comprehensive understanding of the driver's state. Self-attention is applied within each modality to extract refined features, while cross-attention facilitates complementary inter-modality interactions during classification. Comparative evaluations against multiple baseline models demonstrate that the proposed approach achieves state-of-the-art performance in both binary and ternary cognitive load classifications. This framework shows strong potential for enhancing autonomous driving systems through accurate monitoring of drivers' states and timely intervention, as well as advancing the development of Brain–Computer interface systems.

ACKNOWLEDGMENT

This study was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems under Grant 25ZB1200].

REFERENCES

[1] J. Sweller, *Cognitive load theory.*, ser. The psychology of learning and motivation. San Diego, CA, US: Elsevier Academic Press, 2011, pp. 37–76. [Online]. Available: https://doi.org/10.1016/B978-0-12-387691-1.00002-8

[2] V. Nagy, G. Kovács, P. Földesi, D. Kurhan, M. Sysyn, S. Szalai, and S. Fischer, "Testing road vehicle user interfaces concerning the driver's cognitive load," *Infrastructures*, vol. 8, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2412-3811/8/3/49

[3] D. L. Strayer, J. M. Cooper, R. M. Goethe, M. M. McCarty, D. J. Getty, and F. Biondi, "Assessing the visual and cognitive demands of in-vehicle information systems," *Cognitive Research: Principles and Implications*, vol. 4, no. 1, p. 18, Jun 2019. [Online]. Available: https://doi.org/10.1186/s41235-019-0166-3

[4] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2022.

[5] J. Cui, Z. Lan, O. Sourina, and W. Müller-Wittig, "Eeg-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7921–7933, 2023.

[6] D. Pulver, P. Angkan, P. Hungler, and A. Etemad, "Eeg-based cognitive load classification using feature masked autoencoding and emotion transfer learning," 2023. [Online]. Available: https://arxiv.org/abs/2308.00246

[7] Y. Liu, Z. Jia, and H. Wang, "Emotionkd: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 6122–6131. [Online]. Available: https://doi.org/10.1145/3581783.3612277

[8] M. M. Azizi and B. BabaAli, "Biosignals based automated driver cognitive load assessment using a pre-trained transformer," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2024.

[9] R. Qiu and R. Selvan, "Uniphynet: A unified network for multimodal physiological raw signal classification," 2025. [Online]. Available: https://arxiv.org/abs/2507.14163

[10] C. Yang, M. Westover, and J. Sun, "Biot: Biosignal transformer for cross-data learning in the wild," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 78 240–78 260.

[11] P. Angkan, B. Behinaein, Z. Mahmud, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, "Multimodal brain–computer interface for in-vehicle driver cognitive load measurement: Dataset and baselines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5949–5964, 2024.

[12] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018, special issue on deep reinforcement learning. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608017302976

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 3–19.

[15] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] P. Angkan, B. Behinaein, Z. Mahmud, A. Bhatti, D. Rodenburg, P. Hungler, and A. Etemad, "Multimodal brain–computer interface for in-vehicle driver cognitive load measurement: Dataset and baselines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5949–5964, 2024.

[18] F. Farzan, S. Atluri, M. Frehlich, P. Dhami, K. Kleffner, R. Price, S. H. Kennedy *et al.*, "Standardization of electroencephalography for multi-site, multi-platform and multi-investigator studies: Insights from the canadian biomarker integration network in depression," *Scientific Reports*, vol. 7, no. 1, p. 7473, 2017.

[19] M. Carvalho and S. Bras, "Addressing intra-subject variability in electrocardiogram-based biometric systems through a hybrid architecture," *Biomedical Signal Processing and Control*, vol. 87, p. 105465, 2024.