MUSE: Multimodal Uncertainty-Based Self-Driven Evolution for Robust Physiological-Signal–Based Driver Fatigue Detection

Jiaheng Wang [1], Yuan Si[1], Ang Li[1], Zhenyu Wang[1], Tianheng Xu[1], Honglin Hu[21]*

[1]Shanghai Advanced Research Institute
[2]ShanghaiTech University
wangjh@sari.ac.cn, siy@sari.ac.cn, lia@sari.ac.cn, wangzhenyu@sari.ac.cn, xuth@sari.ac.cn, huhl@sari.ac.cn

Abstract

Precise detection of driver mental fatigue is critical for reducing traffic accidents and enhancing road safety. Compared with vision-based detection—which is susceptible to illumination and occlusion—multimodal physiological-signal-based approaches integrate complementary information from diverse biosignals, delivering more faithful and objective fatigue assessments. However, adverse factors such as motion artifacts and environmental noise induce ceaseless deterioration to physiological signals, which markedly degrade the performance of existing multimodal fusion methods. To address this challenge, we propose Multimodal Uncertainty-based Self-driven Evolution, MUSE, reallocating modality contributions in real time via overall uncertainty minimization, thereby enabling efficient collaborative fusion of multi-source predictions. Theoretically, MUSE guarantees a provably bounded cumulative error, and its generalization error approaches the Bayesian-optimal fusion as iterations progress. Operating in a closed loop without labels or manual recalibration, MUSE presents superior suitability for real-world driving scenarios compared to supervised algorithms. On the large-scale driving fatigue dataset SEED-VIG, MUSE outperforms existing models in both classification and regression tasks, substantiating its robustness and practicality as a promising driving fatigue detection solution.

Introduction

Mental fatigue, a central nervous system exhaustion from prolonged high-intensity cognitive load, results in decreased alertness and impaired executive control (Hooda, Joshi, and Shah 2021). Among all contributory factors in traffic accidents, driver fatigue accounts for 8.8–9.5% (Owens et al. 2018), causing substantial casualties and property damage. Critically, empirical evidence shows that alerts issued within one second can reduce such accidents by approximately 90% (Coetzer and Hancke 2011), highlighting the importance of precise and reliable mental fatigue detection.

Fatigue detection methods are broadly classified into vision-based and physiological signal approaches. Vision-based methods leverages cameras to record visual features—such as eye-closure duration, yawning frequency and head-tilt angle—to infer fatigue levels(Lee et al. 2019). Despite significant advances, semantic ambiguity (e.g., open mouths are not always yawns; eyelid closures may signal dryness) and adverse conditions (poor lighting, occlusions, non-frontal camera angles) markedly degrade the precision and robustness of vision-based methods.

In contrast, physiological-signal approaches records electrophysiological data—like electroencephalography (EEG) and electrooculography (EOG)—to quantify driver fatigue. Unlike vision-based methods, biosignals are largely immune to subjective behavior, ambient lighting and camera conditions, making them more faithful indicator of fatigue (Lv et al. 2024). Recognizing that each modality conveys complementary fatigue markers, recent studies integrate multiple physiological modalities to overcome data non-stationarity and information loss in single modality. Specifically, Zhang et al. (Zhang et al. 2023a) apply contrastive learning to align EEG and EOG representations for fatigue regression; Wu et al. (Wu et al. 2020) use a multichannel deep auto-encoder with subnetwork neurons that concatenates latent EEG and EOG embeddings at the fatigue-feature level; and Shi et al. (Shi and Wang 2023) fuse EEG–EOG features via a convolutional auto-encoder (CAE) and feed the combined representation into a recurrent neural network (RNN) for recognition.

While intended to leverage diverse modalities, existing fusion methods fix their parameters after training, preventing adaptation to real signal variability. In practice, EEG and EOG undergo severe fluctuations, due to electrode contact variability, motion artifacts or environmental noise (Wang et al. 2025). For EEG, physiological noise contributes about 11 % of the total power (Scarciglia et al. 2023) and extreme noise can entirely deprive the electrodes of any information (Taleb et al. 2024). Moreover, EOG signal uncertainty arises (Barbara, Camilleri, and Camilleri 2020) from eye movements (Bulling et al. 2010), electrode pressure fluctuations (Bulling, Roggen, and Tröster 2009) and increasing skin impedance (Heide et al. 1999).

Static fusion methods remain rigid to these ongoing signal quality changes. Studies confirm that, when essential features of a certain modality are obscured, static performance plummets, even falling below the best single-modality models (Zhang et al. 2023b) (Zhang et al. 2023b; Wang, Tran, and Feiszli 2020). Furthermore, these empirically inspired

*Corresponding Author.

methods generally deficient in theoretical rigor, tend to converge to inferior optima, and offer limited scope for improvement. Addressing these challenges demands multimodal fusion algorithms capable of dynamically adapting to signal variations while offering rigorous theoretical guarantees. More importantly, given the difficulty of obtaining physiological data labels and human Intervention in driving, operating solely on unlabeled data and performing automatically is imperative (Wang et al. 2024). This explains why label-dependent online algorithms prove impractical in real-world driving scenarios.

Therefore, we propose Multimodal Uncertainty-based Self-driven Evolution, MUSE. By interpreting each modality's uncertainty as a quality indicator and minimizing overall uncertainty, MUSE permits each modality to engage in adaptive, decision-level cooperative voting, thus enabling complementary multi-source fusion. Theoretically, driven solely on readily available unlabeled data, MUSE provides provable bounds on cumulative error and guarantees that its generalization gap asymptotically approaches the Bayesian-optimal fusion. This provides one of the few theoretical analysis of label-free online learning. MUSE's greatest application value is its capacity to operate effectively under label-free driving scenarios, allowing it to exceed a variety of supervised algorithms. In experiments on SEED-VIG, a large multimodal driving fatigue detection dataset with 23 subjects, MUSE outperforms existing fatigue detection models in inter-subject evaluations and maintains the highest performance on new subjects, confirming its robustness and adaptability. The automated evolution of MUSE is also corroborated by the results. Together, MUSE offer a practical approach to fatigue detection with potential to enhance driving safety.

Method

Problem Formulation

Notation In practical application, multimodal signals are acquired at regular intervals, forming a test dataset $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_t, y_t)\}_{t=1}^{T} = \left\{ \left(\{\mathrm{x}_t^{(1)}, \ldots, \mathrm{x}_t^{(m)}, \ldots, \mathrm{x}_t^{(|\mathcal{M}|)}\}, y_t \right) \right\}_{t=1}^{T}$. Here, $\mathbf{x}_t$ aggregates observations from all modalities, and y_t denotes the associated label. $\mathrm{x}_t^{(m)}$ is the data from the m-th modality. The set of modalities is $\mathcal{M}$ with $|\mathcal{M}|$ elements.

Let $\phi^{(m)}$ denote the predictive network for the m-th modality, and $\phi_t^{(m)}$ denote its output at time t. Stacking all $\phi_t^{(m)}$ yields $\boldsymbol{\phi}_t$. In decision-level fusion, $\alpha_t^{(m)}$ is the weight assigned to the m-th modality at time t. All $\alpha_t^{(m)}$ form the weight vector $\boldsymbol{\alpha}_t$. $\boldsymbol{\alpha}_t$ is drawn from the weight set $\mathcal{W}$, a $|\mathcal{M}|$-dimensional probability simplex $\Delta^{|\mathcal{M}|-1}$, defined as $\mathcal{W} = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{M}|} : \alpha^{(m)} \geq 0, \sum_{m=1}^{|\mathcal{M}|} \alpha^{(m)} = 1 \right\}$.

Given $\mathbf{x}_t$, the decision-level fusion output is $\hat{y}_t = \boldsymbol{\alpha}_t^\top \boldsymbol{\phi}_t$, i.e., a decision produced by weighted voting across modalities. $\boldsymbol{\alpha}_t$ critically determines overall performance. Our goal is to continuously update $\boldsymbol{\alpha}_t$ to track modality quality fluctuations and approach the optimal multimodal voting. Considering the difficulty of obtaining labeled signals in driving, the method should be label-free.

Derivation of the Proposed Algorithm

MUSE features a clear derivation: by dynamically adjusting each modality's decision proportions, it minimizes overall predictive uncertainty of the multimodal fatigue detection system, thereby mitigating the influence of signal variability. We use an operator $\mathcal{U}$ to quantify the uncertainty of the m-th modality at time t: $\mathfrak{U}_t^{(m)} = \mathcal{U}\big(\phi^{(m)}, \mathbf{x}_t^{(m)}\big)$. A larger $\mathfrak{U}_t^{(m)}$ indicates reduced signal quality. Effective estimation techniques—such as Monte Carlo Dropout (MC Dropout) and Deep Ensemble—can be adopted; their comparative performance is analyzed in the experiment. Then, we update $\boldsymbol{\alpha}_t$ by solving the optimization:

$$\boldsymbol{\alpha}_{t+1} = \arg\min_{\boldsymbol{\alpha} \in \mathcal{W}} \Big\{ \boldsymbol{\alpha}_t^\top \, \mathfrak{U}_t \; + \; \frac{1}{\eta} \sum_{m=1}^{|\mathcal{M}|} \alpha^{(m)} \ln\!\Big(\frac{\alpha^{(m)}}{\alpha_t^{(m)}} \Big) \Big\}. \quad (1)$$

Here, $\mathrm{KL}\big(\boldsymbol{\alpha} \,\|\, \boldsymbol{\alpha}_t\big) = \sum_{m=1}^{|\mathcal{M}|} \alpha^{(m)} \ln\!\Big(\frac{\alpha^{(m)}}{\alpha_t^{(m)}} \Big)$ denotes the Kullback–Leibler (KL) divergence between $\boldsymbol{\alpha}_{t+1}$ and $\boldsymbol{\alpha}_t$.

$\boldsymbol{\alpha}_t^\top \mathfrak{U}_t$ is the weighted sum of uncertainty over modalities; optimizing it naturally down-weights high-uncertainty modalities and amplifies reliable ones. However, minimizing this term alone pushes all weight onto the lowest-uncertainty modality, creating over-reliance. Moreover, if that modality later falters, its weight can drop from 1 to 0, destabilizing predictions.

To this end, we add a KL divergence regularizer to restrict the change between $\boldsymbol{\alpha}_{t+1}$ and $\boldsymbol{\alpha}_t$. This prevents collapse onto a single modality and allows full exploitation of modalities. η controls the update stability: small η yields conservative updates, whereas a larger η increases sensitivity to the modality quality. With Eq. (1), we derive the weight-update formula as follows. Define the Lagrangian

$$\mathrm{L}_{lag}(\boldsymbol{\alpha}, \lambda) = \boldsymbol{\alpha}_t^T \mathfrak{U}_t + \frac{1}{\eta} \sum_{m=1}^{|\mathcal{M}|} \alpha^{(m)} \ln\!\Big(\frac{\alpha^{(m)}}{\alpha_t^{(m)}} \Big) + \lambda \left(\sum_{m=1}^{|\mathcal{M}|} \alpha^{(m)} - 1 \right), \quad (2)$$

in which the first two terms reproduce the objective in Eq. (1). Taking $\partial \mathrm{L}_{\mathrm{lag}} / \partial \boldsymbol{\alpha} = 0$ yields the update:$\boldsymbol{\alpha}_{t+1} = \frac{\boldsymbol{\alpha}_t \odot \exp(-\eta \mathfrak{U}_t)}{\sum_{m=1}^{|\mathcal{M}|} \alpha_t^{(m)} \exp\left(-\eta \mathfrak{U}_t^{(m)}\right)}$.

Algorithm Description

With the update formula at step t established, we proceed to outline the full iteration. Fig. 1 depicts the MUSE framework. At each time step $t > 1$, we iterate over the following steps:

First, multimodal signals are captured and processed by specialized networks to make predictions $\hat{y}_t^{(m)} = \phi^{(m)}\big(x_t^{(m)}\big)$, and the predictive uncertainty of each modality is estimated by $\mathfrak{U}_t^{(m)} = \mathcal{U}\big(\phi^{(m)}, x_t^{(m)}\big)$.

Then, according to the update formula of $\boldsymbol{\alpha}_t$, the previous voting weights $\alpha_{t-1}^{(m)}$ are multiplied by $e^{-\eta \, \mathfrak{U}_t^{(m)}}$ and the whole weight vector is normalized, acquiring $\boldsymbol{\alpha}_t$.

Finally, the updated weights $\boldsymbol{\alpha}_t$ are used to compute the weighted sum of the modality predictions, yielding the collaborative voting output of step t: $\hat{y}_t = \sum_{m=1}^{|\mathcal{M}|} \alpha_t^{(m)} \, \hat{y}_t^{(m)}$.

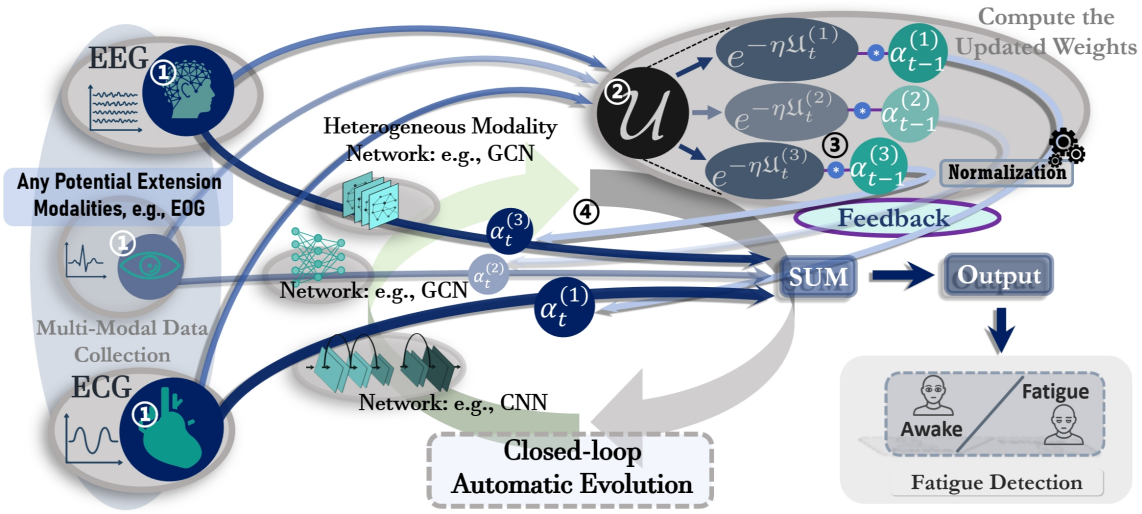

Figure 1: Framework of MUSE

MUSE establishes a human–machine closed-loop system that continuously evolves in non-stationary, label-scarce scenarios and progressively approximates the Bayesian-optimal multimodal fusion strategy.

Our work focuses on efficient multimodal collaborative voting. The individual modal feature extraction network is detailed in the Appendix.

Cumulative Error During Iteration

A central challenge when using unlabeled data is that, without ground-truth, how to control iterative error within a predefined range. In this section, we demonstrate that MUSE enforces a strict upper bound on cumulative error at each update, ensuring that test-set error would not diverge and confirming the iteration's stability and controllability.

First, we define the optimal fusion weights formally. Let $\boldsymbol{\Upsilon}^* = (u^{(1)}, u^{(2)}, \ldots, u^{(|\mathcal{M}|)}) \in \mathcal{W}$ be the unique weight achieving optimal performance, i.e., minimizing the expected loss of the multimodal system on the test set:

$$\boldsymbol{\Upsilon}^* = \arg\min_{\boldsymbol{\alpha}\in\mathcal{W}} \Big\{ \mathbb{E}_{(\mathbf{x}_t, y_t)\sim\mathbb{P}} \big[\sum_{m=1}^{|\mathcal{M}|} \alpha^{(m)} \, \mathcal{L}\big(\phi^{(m)}(\mathbf{x}_t^{(m)}),\, y_t\big) \big] \Big\}, \tag{3}$$

where $\mathbb{P}$ denotes the multimodal data distribution, $\mathcal{L}\big(\phi^{(m)}(\mathbf{x}_t^{(m)}),\, y_t\big)$ denotes the loss incurred by modality m, hereafter abbreviated as $\mathcal{L}_t^{(m)}$. Eq. (3) indicates that $\boldsymbol{\Upsilon}^*$ is the optimal weight that minimizes the total expected loss aggregated over all modalities on the test set. Note that $\boldsymbol{\Upsilon}^*$ is not fixed but adapts to the input modalities to minimize the expected loss, serving as a theoretical construct—the optimal (yet unobservable) set of weights on the test set—used in the regret analysis.

Then, we define R_{muse} as the gap between MUSE's cumulative loss on the test set and that of the optimal fusion after T decisions. As $T \to \infty$, a tighter upper bound on R_{muse} implies that MUSE's cumulative test-set loss converges more closely to the optimal weights. The cumulative loss of MUSE and that of the optimal strategy can be computed as

$$\mathcal{L}_{muse} = \sum_{t=1}^{T} \sum_{m=1}^{|\mathcal{M}|} \alpha_t^{(m)} \mathcal{L}_t^{(m)}, \quad \mathcal{L}_{\boldsymbol{\Upsilon}^*} = \sum_{t=1}^{T} \sum_{m=1}^{|\mathcal{M}|} u^{(m)} \mathcal{L}_t^{(m)}. \tag{4}$$

Then, we present Theorem 1:

Theorem 1 (Cumulative Loss Gap). *Since* $\mathrm{R}_{muse} = \mathcal{L}_{muse} - \mathcal{L}_{\boldsymbol{\Upsilon}^*}$, *then we have*

$$\mathrm{R}_{muse} \le \left(\frac{\lambda_2}{\lambda_1} - 1 \right) \mathcal{L}_{\boldsymbol{\Upsilon}^*} + \frac{\mathcal{G}\sqrt{2T \ln |\mathcal{M}|}}{\lambda_1}, \tag{5}$$

assuming there exist constants $\lambda_1, \lambda_2 > 0$ *with* $\lambda_1 \le \lambda_2$ *such that* $\mathfrak{U}_t^{(m)}/\lambda_2 \le \mathcal{L}_t^{(m)} \le \mathfrak{U}_t^{(m)}/\lambda_1$, *i.e. the loss does not diverge to infinity.* $\mathcal{G}$ *denotes the upper bound on modality uncertainty* ($\mathcal{G} < 1$).

Eq. (5) bounds MUSE's cumulative loss over T iterations relative to the optimal fusion. This implies that, during evolution, MUSE's predictive error is always maintained within a controlled range.

Performance Comparison with Optimal Fusion After Iterations

Next, we compare MUSE's and the optimal fusion's expected predictive error (expected risks) on future samples after T iterations. By contrast, Eq. (5) highlights overall performance during adaptation (including its initial phase). We show that, as T grows, MUSE's performance approaches the optimal fusion, confirming its continuous evolution.

For any fusion weights $w \in \mathcal{W}$, define its expected risks as $\mathcal{E}_w = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathcal{L}(w; x, y)\big]$. Let $\mathcal{E}_{\text{muse}}$ and $\mathcal{E}_{\boldsymbol{\Upsilon}^*}$ denote the expected risks of MUSE and the optimal fusion respectively. Here, the expected risks of MUSE refers to the expected predictive error of MUSE's voting weights after T iterations on future samples—i.e., the post-adaptation accuracy. Applying Eq. (5) yields:

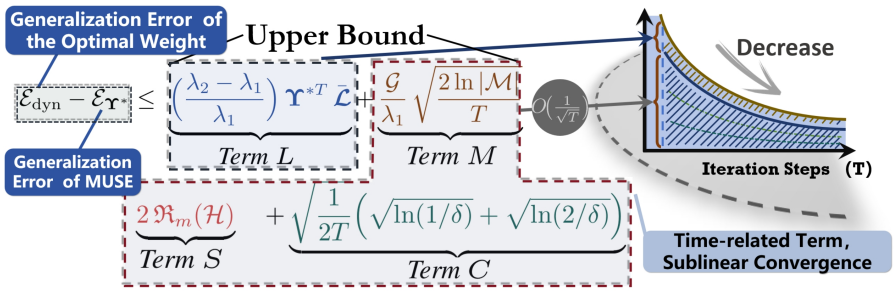

Figure 2: depicts the upper bound on the expected risk gap between MUSE and the optimal fusion, comprising four Terms (L, M, S, C). $Term\, L, S$, and C together constitute the time-related terms. The curves on the right accentuate the individual contributions of each term. Notably, time-related terms decays at a rate of $O(1/T)$ as the number of iterations increases.

Theorem 2 (Generalization Gap). *With probability at least* $1-2\delta$,

$$\mathcal{E}_{\text{muse}} - \mathcal{E}_{\boldsymbol{\Upsilon}^*} \leq \underbrace{\left(\tfrac{\lambda_2-\lambda_1}{\lambda_1}\right)\boldsymbol{\Upsilon}^{*T}\bar{\boldsymbol{\mathcal{L}}}}_{Term\ L} + \underbrace{\tfrac{\mathcal{G}}{\lambda_1}\sqrt{\tfrac{2\ln|\mathcal{M}|}{T}}}_{Term\ M} + \underbrace{2\,\mathfrak{R}_m(\mathcal{H})}_{Term\ S} + \underbrace{\sqrt{\tfrac{1}{2T}}\left(\sqrt{\ln(1/\delta)}+\sqrt{\ln(2/\delta)}\right)}_{Term\ C}. \quad (6)$$

Here, $\bar{\boldsymbol{\mathcal{L}}} = \left(\bar{\mathcal{L}}^{(1)}, \ldots, \bar{\mathcal{L}}^{(m)}, \ldots, \bar{\mathcal{L}}^{(|\mathcal{M}|)}\right)^T$ is the vector of mean losses, where $\bar{\mathcal{L}}^{(m)} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t^{(m)}$ denotes the average loss of the m-th modality over T time steps. The inner product $\boldsymbol{\Upsilon}^{*T}\bar{\boldsymbol{\mathcal{L}}}$ therefore measures the cumulative loss of the optimal weights, and $\mathfrak{R}_m(\mathcal{H})$ denotes the Rademacher complexity of the hypothesis class $\mathcal{H}$ ($\mathcal{H}$ comprises all functions mapping multimodal input $\{\mathbf{x}_t^{(m)}\}_{m=1}^{|\mathcal{M}|}$ to the fused output $\hat{y}_t$).

Eq. (6) presents the upper bound on the gap in generalization error between MUSE and the optimal weights. An intuitive illustration of Eq. (6) is provided in Fig. 2. We explain the four terms in Theorem 2: **Term L (Loss function fluctuation Term)** quantifies the impact of variability in the per-step loss on the generalization gap. Next, **Term S (Sparsity Regularization Term)** measures the contribution of model complexity to the generalization gap. **Term C (Confidence Interval Term)** scales inversely with the confidence level δ (it quantifies that Eq. (6) holds with probability at least $1-\delta$). Finally, **Term M (Modality Uncertainty Term)** quantifies the effect of the uncertainty upper bound $\mathcal{G}$ on generalization error. Analyzing Theorem 2 yields the Lemma 1.

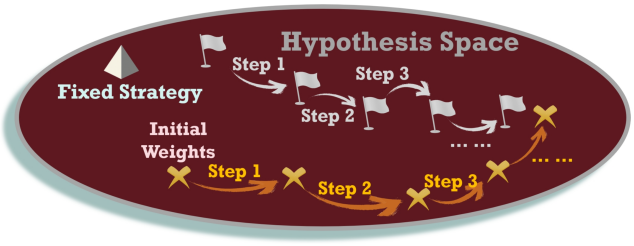

Figure 3: Iterative Convergence of MUSE to the Optimal Fusion

Lemma 3 (Asymptotic Behavior of MUSE). *Let* $\Omega_{muse} = \mathcal{E}_{muse} - \mathcal{E}_{\boldsymbol{\Upsilon}^*}$. *Consider the asymptotic regime as* $T \to \infty$; *at this limit,* $Term\, S, M$, *and* C *vanish, removing their contributions to* Ω_{muse} *and causing MUSE's expected risk to approach that of the optimal fusion.*

Lemma 1 implies that, although MUSE may initially diverge from the optimal multimodal voting, as iterations progress, $Term\, S, M$, and C decrease at a rate of $O\left(1/\sqrt{T}\right)$ and ultimately converge to zero. This markedly reduces the long-run performance gap between MUSE and the optimal voting, enabling automated, label-free adaptation to highly dynamic signals.

Moreover, once T is sufficiently large, Ω_{muse} is determined solely by $Term\, L$, which in turn is driven by the gap between the loss bounds: the lower bound fixes λ_2 and the upper bound fixes λ_1. In a well-trained model or under specialized training regimes where losses stabilize such that $\lambda_2 \approx \lambda_1$, $Term\, L$ vanishes. In that case, MUSE can achieve virtually identical performance to the optimal fusion.

Fig. 3 illustrates MUSE's iterative convergence to the optimal fusion. The brown ellipse delineates the hypothesis space of feasible weights, the tetrahedron denotes the fixed weights obtained on the training set, and the white flag marks the optimal fusion on the test set. Crosses show MUSE's successive updates, following the arrows as they progressively approach the optimal fusion.

Experiments

Dataset and Data Preprocessing

SEED-VIG (Zheng and Lu 2017) is an open large dataset for driving vigilance research with recordings from 23 subjects to ensure sufficient sample size. The experiment employs a simulated driving task that closely mimics real-world, lasts about two hours. The experiment employed a genuine cockpit setup—comprising a seat and steering wheel facing a large LCD that displayed a four-lane highway in real time. Accelerator and steering inputs were processed to update the scene synchronously. A straight, monotonous road segment was chosen, and trials were conducted in the early postprandial period—when somnolence peaks—to maximize fatigue induction. EEG was acquired via the 10–20 system using 17 channels (see Fig. 4 left), alongside four-channel EOG (see

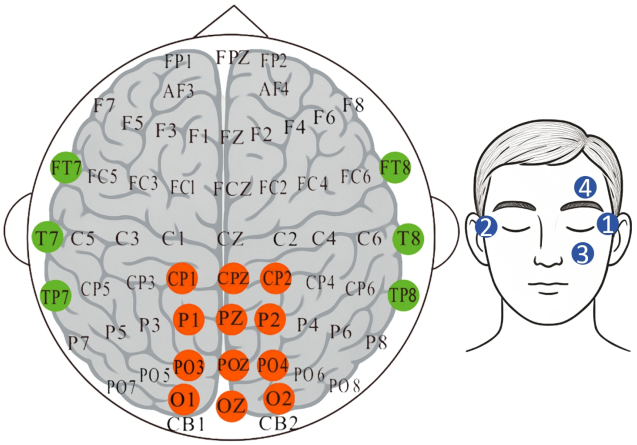

Figure 4: Spatial Distribution of Electrodes

Fig. 4 right). Both EEG and EOG were sampled at 1000 Hz, down-sampled to 200 Hz, and band-pass filtered between 1–75 Hz to suppress noise.

In EEG preprocessing, differential entropy (DE) is adopted as a vigilance feature. Assuming the raw signal follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, its differential entropy is defined as $DE(EEG) = \frac{1}{2}\log_2\left(2\pi e\,\sigma^2\right)$.

Using Eq. (7), we compute DE for five bands—δ (1–4 Hz), θ (4–8 Hz), α (8–14 Hz), β (14–31 Hz), and γ (31–50 Hz)—as follows. The continuous EEG is first partitioned into non-overlapping 8 s windows. For each window, we perform a short-time Fourier transform to estimate the power spectral density within each band, summing the band power to obtain the estimate of σ^2. Substituting this estimate into Eq. (7) yields the band-specific DE values for every channel.

Raw EOG is recorded in two channels: horizontal (HEO) for saccadic movements and vertical (VEO) for blinks. To enable saccade and blink detection using the algorithm proposed in (Zheng and Lu 2017), continuous wavelet coefficients were obtained via the Mexican hat wavelet at scale 8 as $\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}}\left(1 - \frac{t^2}{\sigma^2}\right)e^{-t^2/(2\sigma^2)}$.

Here, σ is the standard deviation. Within each non-overlapping 8 s window, 36 features are extracted from the detected eye events, as detailed in the Appendix. This feature set and extraction pipeline match those described in (Zheng and Lu 2017).

SEED-VIG employs the percentage of eye closure (PERCLOS) as the groud-truth of fatigue levels for each 8 s sample, defined by $\text{PERCLOS} = \frac{\text{eye\_closing\_time}}{\text{total\_time}}$. Here, eye-closing time is obtained from SMI eye-tracking glasses. Windows with PERCLOS exceeding 30% are labeled "fatigued," and others "alert," yielding binary labeling.

Evaluation Metrics and Experimental Setup

For a comprehensive evaluation, we employ both regression metrics on the continuous PERCLOS outputs and classification metrics by thresholding the outputs into "fatigued" versus "alert" classes. Specifically, regression performance is quantified by mean square error (RMSE)—measuring average prediction error—and Pearson's correlation coefficient (CORR)—assessing the agreement between predicted and actual fatigue trends. For a set of N samples, these metrics are defined as $\text{RMSE}(Y, \hat{Y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$, and $\text{CORR}(Y, \hat{Y}) = \frac{\sum_{i=1}^{N}(y_i - \bar{y})\,(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\,\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}$, where $Y = (y_1, \ldots, y_N)$ and $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_N)$ are the vectors of true labels and model predictions, respectively, with $\bar{y}$ and $\bar{\hat{y}}$ denoting their sample means. In general, higher CORR and lower RMSE correspond to better model performance.

For the binary classification, we employ accuracy, precision, recall, and F1 score to provide a comprehensive assessment of performance.

For evaluation, we conduct intra-subject and inter-subject experiments: for intra-subject, each subject's same-day data are split into five contiguous folds (4 for training, 1 for testing); for inter-subject, one subject is left out for testing while the rest for training. In both settings, we report RMSE and CORR averaged over all subjects. During testing, each 8 s sample defines a time step t. At each step, MUSE processes multimodal input $\mathbf{x}_t$ to generate the prediction $\hat{y}_t$, which is then compared with the label y_t. All experiments were conducted on an RTX 3070 Ti GPU (8 GB VRAM). Initially, η_0 was set to 1.17 via a validation set comprising one-eighth of the training data. During the iteration, η decays proportionally to $\eta_0/\sqrt{T}$.

For uncertainty estimation, we adopt MC Dropout (Gal and Ghahramani 2016), which interprets dropout as a Bayesian approximation by maintaining stochastic neuron deactivations at inference. Gal and Ghahramani (Gal and Ghahramani 2016) proved that MC Dropout captures model uncertainty. In MC Dropout, dropout layers remain active at test time, producing varied network paths. At each time step t for m-th modality with input $x_t^{(m)}$, we perform R stochastic forward passes, yielding a set of predictions

$$\{\hat{y}_{t,r}^{(m)}\}_{r=1}^{R}, \qquad \hat{y}_{t,r}^{(m)} = f_{\text{drop}}^{(m)}\left(x_t^{(m)}\right), \quad r = 1, \ldots, R. \tag{7}$$

Use the variance of predictions to quantify uncertainty:

$$\bar{\hat{y}}_t^{(m)} = \frac{1}{R}\sum_{r=1}^{R}\hat{y}_{t,r}^{(m)}, \quad \mathfrak{U}_t^{(m)} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{y}_{t,r}^{(m)}\right)^2 - \left(\bar{\hat{y}}_t^{(m)}\right)^2. \tag{8}$$

Results and Analysis

Prediction Performance Table 1 summarizes the inter-subject classification performance on SEED-VIG for all

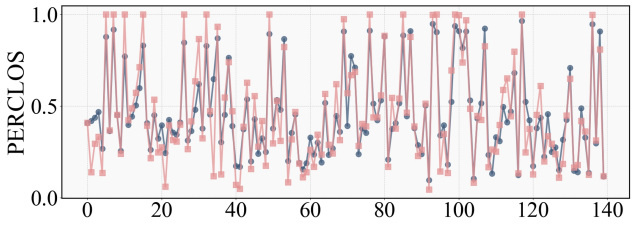

Figure 5: Predicted PERCLOS vs. Ground Truth

Method	Accuracy	Precision	Recall	F1-score
EEGNet	72.7	74.0	87.9	80.4
AMS-CNN	80.2	82.0	87.4	84.4
ST-Encoding	70.0	77.1	72.6	74.8
CNN-Attention	78.6	75.4	76.5	75.9
3DCNN-LSTM	75.7	78.3	76.4	77.3
MMFNet	69.6	64.0	69.2	66.5
SM model	89.8	88.6	91.1	89.8
ICNN	79.6	75.2	78.5	76.8
SFT-Net	87.1	88.2	92.2	90.1
Conformer-O	81.7	83.8	77.3	80.2
Conformer ω BN	88.4	87.1	89.6	88.3
CSF-GTNet	81.5	76.2	80.4	78.2
CM-FusionNet	84.6	85.4	85.5	85.3
MUSE (ours)	**92.2**	**91.6**	**92.5**	**91.3**

Table 1: Inter-subject Classification Performance Comparison of Various Methods on SEED-VIG

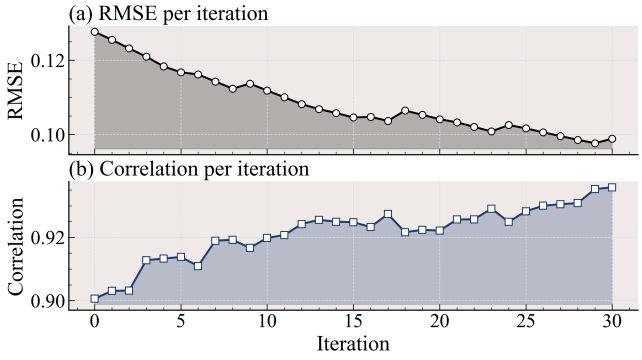

Figure 6: MUSE Evolution with MC Dropout

compared fatigue-detection methods—EEGNet (Lawhern et al. 2018), AMS-CNN (Gao et al. 2021), ST-Encoding (Paulo, Pires, and Nunes 2021), CNN-Attention (Xu et al. 2021), 3DCNN-LSTM (Wu et al. 2021), MMFNet (Zhang et al. 2021), SM model (Li, Wang, and Sourina 2022), ICNN (Cui et al. 2022), SFT-Net (Gao et al. 2023), Conformer-O (Song et al. 2023), Conformer,ω,BN (Song et al. 2023), CSF-GTNet (Gao et al. 2024), CM-FusionNet (Huang et al. 2025)—and the proposed MUSE framework. MUSE achieves the highest Accuracy (92.2%), Precision (91.6%), Recall (92.5%), and F1-score (91.3%), evidencing its superior robustness to inter-subject variability and signal-distribution shifts. By contrast, unimodal methods such as EEGNet, AMS-CNN, and 3DCNN-LSTM—limited to EEG and neglecting the complementary information from EOG—deliver lower performance.

Fig. 5 shows MUSE's PERCLOS tracking over continuous 140 samples for a certain subject. The panel plots sample index (x-axis) versus PERCLOS (y-axis), with MUSE's predictions in blue and the ground-truth in translucent red. The two curves closely overlap, demonstrating MUSE's rapid adaptation and precise fatigue estimation.

Method	RMSE	CORR	Method	RMSE	CORR
ELM	0.11	0.78	MCDAEsn	0.11	0.76
AE-ELM	0.10	0.82	CAE-LSTM	0.09	0.95
SVR	0.10	0.83	CAE-BiLSTM	0.08	0.96
CCRF	0.10	0.84	MVE-MPCL	0.09	0.89
CCNF	0.09	0.85	MSCNN	0.08	0.91
DAE	0.09	0.85	MSCNN-CAM	0.07	0.93
DNNSN	0.08	0.86	**MUSE (ours)**	**0.06**	**0.97**

Table 2: Regression Performance Comparison of MUSE and other benchmarks

Table 2 summarizes intra-subject regression performance (RMSE and CORR) on SEED-VIG. A range of static fusion methods—ELM (Huang et al. 2012), AE-ELM (Yang and Wu 2016), SVR (Tian, Zhu, and et al. 2014), CCRF (Baltrusaitis, Banda, and Robinson 2013), CCNF (Imbrasaite, Baltrusaitis, and Robinson 2014), DAE (Du, Liu, and et al. 2017), DNNSN (Wu et al. 2018), MCDAEsn (Wu et al. 2020), CAE-LSTM (Shi and Wang 2023), CAE-BiLSTM (Shi and Wang 2023), and MVE-MPCL (Zhang et al. 2023a)—yield RMSE between 0.08–0.11 and CORR from 0.76–0.96. By contrast, MUSE achieves RMSE of 0.06 and CORR of 0.97, outperforming all static approaches. This highlights MUSE's adaptability to signal variability and its capacity for more precise fatigue detection.

Method	Intra-subject		Inter-subject	
	RMSE	CORR	RMSE	CORR
CONCAT	0.079	0.901	0.129	0.813
ETF	0.070	0.929	0.119	0.838
DCCA	0.078	0.917	0.125	0.816
DynMM	0.066	0.943	0.115	0.846
PDF	0.067	0.939	0.116	0.841
QMF	0.063	0.945	0.105	0.855
EEG	0.087	0.897	0.132	0.810
EOG	0.076	0.907	0.123	0.830
MUSE (ours)	**0.060**	**0.966**	**0.097**	**0.866**

Table 3: Comparison of Multimodal Fusion Strategies

Table 3 compares the performance of static fusion methods (CONCAT (Ngiam et al. 2011), DCCA (Andrew et al. 2013)), recent dynamic fusion strategies (ETF (Wang et al. 2021), DynMM (Xue and Marculescu 2023), PDF (Cao et al. 2024), QMF (Zhang et al. 2023b)) and unimodal baselines (EEG, EOG) under identical feature-extraction backbones, providing a controlled, fair evaluation of multimodal fusion strategies. In the intra-subject setting, MUSE achieves the lowest RMSE of 0.060 and the highest CORR of 0.966. By contrast, static fusion methods yield RMSE in the range 0.078–0.079 and CORR in 0.901–0.917, while dynamic fusion methods fall in RMSE 0.063–0.070 and CORR 0.929–0.945. The unimodal EEG and EOG baselines score RMSE 0.087/0.076 and CORR 0.897/0.907, respectively. In the more challenging inter-subject setting, MUSE again at-

tains the best results with RMSE 0.097 and CORR 0.866, outperforming static fusion (RMSE 0.125–0.129, CORR 0.813–0.816), dynamic fusion (RMSE 0.105–0.119, CORR 0.838–0.846) and unimodal baselines (RMSE 0.132/0.123, CORR 0.810/0.830). These results yields two conclusions: dynamic fusion consistently surpasses both static fusion and unimodal baselines; and MUSE achieves optimal performance in both experimental settings, validating its efficacy and robustness.

The computational efficiency of MUSE mainly depends on the uncertainty estimation strategy. The heteroscedastic regression version is light and avoids iterative sampling, giving much lower latency and computation (average 1.62 ms, 2.87 M FLOPs per sample, manageable on embedded hardware) than MC Dropout (90.31 ms, 160.95 M FLOPs). This shows that MUSE's feasibility for practical deployment stems from an efficient uncertainty estimator.

Experimental Insights Guiding MUSE's Design This part aims to provide an intuitive design rationale for MUSE by contrasting the spatial distribution of DE power between low- and high-uncertainty samples. Fig. 7 shows α-band DE power topographies for samples with high fatigue (PERCLOS 0.65–0.95), comparing six low-uncertainty cases (uncertainty $<$ 0.16; uncertainties from left to right: 0.134, 0.128, 0.156, 0.112, 0.121, 0.140) and six high-uncertainty cases (uncertainty $>$ 0.60; uncertainties from left to right: 0.752, 0.810, 0.749, 0.863, 0.662, 0.617). Colors span from blue (lowest DE power) to orange (highest DE power). As shown in (Zheng and Lu 2017), high-fatigue states evoke pronounced α-band DE power elevations over posterior cortices—particularly the temporal and parietal lobes—a key physiological marker for fatigue detection. In the low-uncertainty maps, clear posterior (temporal/parietal) α-band hotspots appear, reflecting the well-established fatigue biomarker; the high-uncertainty samples, by contrast, exhibit diffuse, heterogeneous DE distributions that obscure the expected signature. This pattern is also clear in the full dataset, supported by a statistical test in the appendix.

This inform MUSE's design: when a modality's uncertainty is low and its fatigue features are distinct, MUSE increases its weight to exploit reliable information; when uncertainty is high and the biomarker is masked, MUSE reduces the weight to maintain overall stability.

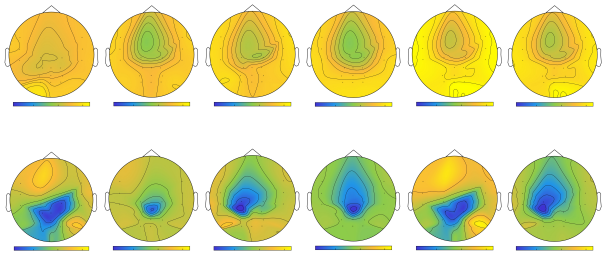

Figure 7: Scalp topography maps of EEG DE for low-uncertainty samples (top row) and high-uncertainty samples (bottom row).

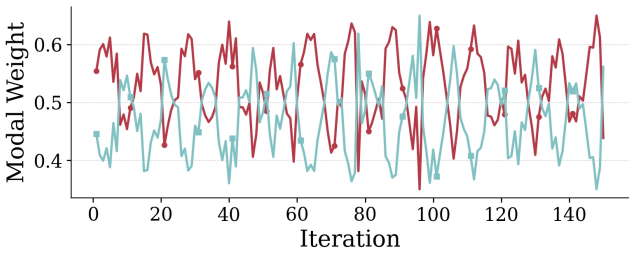

Figure 8: Modal Weight Dynamics over Iterations

Automatic Evolution of MUSE Now we evaluate MUSE's iterative convergence when instantiated with four mainstream uncertainty estimators (Gawlikowski et al. 2023): Bayesian-based methods, Deterministic network methods (Kendall and Gal 2017), ensemble modeling (Lakshminarayanan, Pritzel, and Blundell 2017), and input perturbation (Ayhan and Berens 2018). Fig. 6 plots RMSE and CORR under MC Dropout (a representative Bayesian-based methods): fusion weights are updated using unlabeled data, with performance metrics computed every five updates. Here, η decays with each iteration: at iteration T, it is set to its initial value multiplied by $1/T$. Despite minor fluctuations, RMSE steadily decreases and CORR steadily rises, reaching optimal values at the end of the iteration. The KL-divergence regularizer is crucial in this process, as it constrains abrupt weight shifts, yielding smooth and stable convergence. The evolution plots for the other three uncertainty estimation methods are provided in the Appendix; they also show a uniform RMSE decline and CORR rise. These results underscore MUSE's adaptability to diverse uncertainty estimators and its ability to self-evolve in unlabeled, online multimodal fatigue-detection scenarios. Moreover, as shown in Fig. 8, the weights of EEG (red line) and EOG fluctuate over iterations, indicating the model's adaptive adjustment.

Conclusion

In physiological-signal–based driver fatigue detection, we introduce MUSE, a framework for perpetual adaptation to non-stationary physiological-signals. MUSE reassigns per-modality decision proportions to minimize overall uncertainty, thereby achieving more effective complementary multi-source fusion. Even without any labels, MUSE strictly bounds cumulative error across iterations and automatically approach the optimal fusion strategy. Compared to supervised approaches, MUSE requires neither manual annotations nor adjustment, rendering it well-suited to real-world driving scenarios lacking label availability. On the large-scale dataset SEED-VIG, MUSE outperforms existing physiological-signal–based models, as well as static and dynamic fusion methods. These results show MUSE's robustness to signal-quality variability and its practical promise for driver fatigue detection and road safety improvement. Future work can focus on applying MUSE to real-road cases—with vibration, rich scenes, diverse driver actions, and lighter sensors (e.g., wrist PPG/SpO_2 bands, few-channel dry EEG/EOG, slim ECG, or no-wear eye

tracker)—as this is central to improving practicality.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project under Grant 2024ZD0529206; in part by the National Natural Science Foundation of China under Grant U24A20209 and 62401552; in part by the Shanghai Municipal Commission of Economy and Information Project under Grant 2024-GZL-RGZN-01027. The experiments of this work were supported by the Core Facility Platform of Computer Science and Communication, SIST, ShanghaiTech University.

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255. PMLR.

Ayhan, M. S.; and Berens, P. 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*.

Baltrusaitis, T.; Banda, N.; and Robinson, P. 2013. Dimensional Affect Recognition Using Continuous Conditional Random Fields. In *IEEE International Conference & Workshops on Automatic Face & Gesture Recognition*, 1–8.

Barbara, N.; Camilleri, T. A.; and Camilleri, K. P. 2020. A comparison of EOG baseline drift mitigation techniques. *Biomedical Signal Processing and Control*, 57: 101738.

Bulling, A.; Roggen, D.; and Tröster, G. 2009. Wearable EOG goggles: Seamless sensing and context-awareness in everyday environments. *Journal of Ambient Intelligence and Smart Environments*, 1(2): 157–171.

Bulling, A.; Ward, J. A.; Gellersen, H.; and Tröster, G. 2010. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence*, 33(4): 741–753.

Cao, B.; Xia, Y.; Ding, Y.; Zhang, C.; and Hu, Q. 2024. Predictive Dynamic Fusion. *arXiv preprint arXiv:2406.04802*.

Coetzer, R. C.; and Hancke, G. P. 2011. Eye detection for a real-time vehicle driver fatigue monitoring system. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, 66–71. IEEE.

Cui, J.; Lan, Z.; Sourina, O.; and Müller-Wittig, W. 2022. EEG-Based Cross-Subject Driver Drowsiness Recognition with an Interpretable Convolutional Neural Network. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10): 7921–7933.

Du, L.-H.; Liu, W.; and et al. 2017. Detecting Driving Fatigue with Multimodal Deep Learning. In *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, 74–77.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gao, D.; Li, P.; Wang, M.; Liang, Y.; Liu, S.; Zhou, J.; Wang, L.; and Zhang, Y. 2024. CSF-GTNet: A Novel Multi-Dimensional Feature Fusion Network Based on ConvNeXt-GeLU-BiLSTM for EEG-Signals-Enabled Fatigue Driving Detection. *IEEE Journal of Biomedical and Health Informatics*, 28(5): 2558–2568.

Gao, D.; Wang, K.; Wang, M.; Zhou, J.; and Zhang, Y. 2023. SFT-net: A network for detecting fatigue from EEG signals by combining 4d feature flow and attention mechanism. *IEEE Journal of Biomedical and Health Informatics*.

Gao, Z.; Sun, X.; Liu, M.; Dang, W.; Ma, C.; and Chen, G. 2021. Attention-based parallel multiscale convolutional neural network for visual evoked potentials EEG classification. *IEEE Journal of Biomedical and Health Informatics*, 25(8): 2887–2894.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589.

Heide, W.; König, E.; Trillenberg, P.; Kömpf, D.; and Zee, D. S. 1999. Electrooculography: technical standards and applications. The International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*, 52: 223–240.

Hooda, R.; Joshi, V.; and Shah, M. 2021. A comprehensive review of approaches to detect fatigue using machine learning techniques. *Chronic Diseases and Translational Medicine*.

Huang, F.; Yang, C.; Weng, W.; Chen, Z.; and Zhang, Z. 2025. CM-FusionNet: A cross-modal fusion fatigue detection method based on electroencephalogram and electrooculogram. *Computers and Electrical Engineering*, 123: 110204.

Huang, G.-B.; Zhou, H.; Ding, X.; and Zhang, R. 2012. Extreme Learning Machine for Regression and Multiclass Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2): 513–529.

Imbrasaite, V.; Baltrusaitis, T.; and Robinson, P. 2014. CCNF for Continuous Emotion Tracking in Music. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 1–6.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5): 056013.

Lee, K. H.; Kim, W.; Choi, H. K.; and Jang, B. T. 2019. A study on feature extraction methods used to estimate a

driver's level of drowsiness. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*, 710–713. IEEE.

Li, R.; Wang, L.; and Sourina, O. 2022. Subject matching for cross-subject EEG-based recognition of driver states related to situation awareness. *Methods*, 202: 136–143.

Lv, X.; Zheng, G.; Zhai, H.; Zhou, K.; and Zhang, W. 2024. Driver fatigue detection method based on temporal–spatial adaptive networks and adaptive temporal fusion module. *Computers and Electrical Engineering*, 119: 109540.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Y.; et al. 2011. Multimodal deep learning. In *ICML*, volume 11, 689–696.

Owens, J.; Dingus, T.; Guo, F.; Fang, Y.; Perez, M.; McClafferty, J.; and Tefft, B. 2018. Prevalence of drowsy-driving crashes: Estimates from a large-scale naturalistic driving study.

Paulo, J. R.; Pires, G.; and Nunes, U. J. 2021. Cross-Subject Zero Calibration Driver's Drowsiness Detection: Exploring Spatiotemporal Image Encoding of EEG Signals for Convolutional Neural Network Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29(5): 905–915.

Scarciglia, A.; Catrambone, V.; Bonanno, C.; and Valenza, G. 2023. Physiological noise: Definition, estimation, and characterization in complex biomedical signals. *IEEE Transactions on Biomedical Engineering*, 71(1): 45–55.

Shi, J.; and Wang, K. 2023. Fatigue driving detection method based on Time-Space-Frequency features of multimodal signals. *Biomedical Signal Processing and Control*, 84: 104744.

Song, Y.; Zheng, Q.; Liu, B.; and Gao, X. 2023. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 710–719.

Taleb, F.; Vasco, M.; Rajabi, N.; Björkman, M.; and Kragic, D. 2024. Challenging Deep Learning Methods for EEG Signal Denoising under Data Corruption. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. IEEE.

Tian, Y.; Zhu, Z.; and et al. 2014. Nonparallel Support Vector Machines for Pattern Classification. *IEEE Transactions on Cybernetics*, 44(7): 1067–1079.

Wang, J.; Wang, Z.; Xu, T.; Li, A.; Si, Y.; Zhou, T.; Zhao, X.; and Hu, H. 2025. Enhancing the Reliability of Affective Brain-Computer Interfaces by Using Specifically Designed Confidence Estimator. *IEEE Journal of Biomedical and Health Informatics*.

Wang, J.; Wang, Z.; Xu, T.; Zhou, T.; Zhao, X.; and Hu, H. 2024. SS-MSDA: Streamlined Sample-level Multi-source Domain Adaptation for EEG Emotion Recognition. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. IEEE.

Wang, W.; Tran, D.; and Feiszli, M. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12695–12705.

Wang, Y.; Jiang, W.-B.; Li, R.; and Lu, B.-L. 2021. Emotion transformer fusion: Complementary representation properties of EEG and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1575–1578. IEEE.

Wu, E. Q.; Xiong, P.; Tang, Z.-R.; Li, G.-J.; Song, A.; and Zhu, L.-M. 2021. Detecting Dynamic Behavior of Brain Fatigue Through 3-D-CNN-LSTM. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1): 1–11.

Wu, W.; Sun, W.; Wu, Q. J.; Yang, Y.; Zhang, H.; Zheng, W.-L.; and Lu, B.-L. 2020. Multimodal vigilance estimation using deep learning. *IEEE Transactions on Cybernetics*, 52(5): 3097–3110.

Wu, W.; Wu, Q. J.; Sun, W.; Yang, Y.; Yuan, X.; Zheng, W.-L.; and Lu, B.-L. 2018. A regression method with subnetwork neurons for vigilance estimation using EOG and EEG. *IEEE Transactions on Cognitive and Developmental Systems*, 13(1): 209–222.

Xu, T.; Wang, H.; Lu, G.; Wan, F.; Deng, M.; Qi, P.; Bezerianos, A.; Guan, C.; and Sun, Y. 2021. E-Key: An EEG-Based Biometric Authentication and Driving Fatigue Detection System. *IEEE Transactions on Affective Computing*, 14(2): 864–877.

Xue, Z.; and Marculescu, R. 2023. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2575–2584.

Yang, Y.; and Wu, Q. J. 2016. Multilayer Extreme Learning Machine with Subnetwork Nodes for Representation Learning. *IEEE Transactions on Cybernetics*, 46(11): 2570–2583.

Zhang, M.; Luo, Z.; Xie, L.; Liu, T.; Yan, Y.; Yao, D.; Zhao, S.; and Yin, E. 2023a. Multimodal vigilance estimation with modality-pairwise contrastive loss. *IEEE Transactions on Biomedical Engineering*, 71(4): 1139–1150.

Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J. T.; and Peng, X. 2023b. Provable Dynamic Fusion for Low-Quality Multimodal Data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 1–17. Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

Zhang, Y.; Chen, S.; Cao, W.; Guo, P.; Gao, D.; and Wang, M. 2021. MFFNet: Multi-Dimensional Feature Fusion Network Based on Attention Mechanism for sEMG Analysis to Detect Muscle Fatigue. *Expert Systems with Applications*, 185: 115639.

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