KWS-TA-CNN Network: Towards Lightweight Mild Cognitive Impairment Detection Using Eye-Tracking Signals From Virtual Reality Stroop Test

Menglan Ruan[1], Wenyuan Li[2], Lei Jin[1], Leqi Yang[1], Wenbin Luo[1], Bin Liu[2], Chunfeng Yang [1*](chunfeng.yang@seu.edu.cn), Wentao Xiang [2*](xiangbmu@njmu.edu.cn),

[1]School of Computer Science and Technology, Southeast University, Nanjing, 211189, China.

[2]School of Biomedical Engineering, Nanjing Medical University, Nanjing, 211166, China.

Abstract

Mild cognitive impairment (MCI) detection using eye-tracking (ET) signals in virtual reality (VR)-based cognitive tasks shows great promise, as it can capture rich temporal and behavioral information. Therefore, we build four VR-based tasks based on Stroop test and construct a dataset for MCI detection using ET signals. However, ET signals often suffer from non-stationarity, variability, and redundancy, challenging accurate MCI detection. To address these issues, we propose a novel lightweight network KWS-TA-CNN with three key components: 1) Kymatio Wavelet scattering transform (KWS), which generates time-robust features and reduces memory usage through a depth-first traversal strategy; 2) Temporal Attention (TA) to dynamically weight critical time steps for MCI detection; and 3) 1D Convolutional Neural Network (CNN) to capture local temporal patterns and reduce feature redundancy. Experimental results from leave-one-subject-out cross-validation show high performance, with subject-level accuracies of 0.8158, 0.9211, 0.8158, and 0.8421 across the four tasks, demonstrating its strong clinical potential.

Keywords: MCI detection, eye-tracking signals, Wavelet scattering transform, convolutional neural network, attention

Introduction

Mild Cognitive Impairment (MCI) (Gauthier et al., 2006), an early stage of cognitive decline, is a critical precursor to Alzheimer's disease (AD), where the latter is an irreversible neurodegenerative disorder (Scheltens et al., 2016). Since MCI represents a transitional phase where intervention can still modify the disease's trajectory, early detection, and screening are crucial for preventing or delaying progression to AD (Albert et al., 2013). Despite increasing research into MCI, public awareness of this condition remains limited. This lack of awareness not only hampers public support for MCI research but also obstructs the promotion of early detection and intervention strategies.

The detection of MCI relies on various assessment methods, including psychological evaluations (such as the Montreal Cognitive Assessment, MoCA scale (Nasreddine et al., 2005)), cerebrospinal fluid analyses (Ritchie et al., 2014), neurological examinations, and advanced imaging techniques such as magnetic resonance imaging (Mirzaei & Adeli, 2022; Mofrad & et al., 2021), positron emission tomography (Arbizu, Festari, Altomare, et al., 2018), and electroencephalography (Siuly et al., 2020). However, they come with some certain limitations. Psychological evaluations often require considerable time and expertise to administer and interpret, limiting their scalability and accessibility in clinical or community settings. Invasive procedures like cerebrospinal fluid analysis can cause discomfort for patients, and imaging techniques are expensive and necessitate specialized equipment, further hindering their widespread use.

Virtual reality (VR) technology has gained significant attention in recent years due to its potential in cognitive assessment. VR offers a multi-sensory, highly immersive, and interactive environment through realistic three-dimensional scenarios for conducting cognitive evaluations. These advantages not only enhance patient engagement and interest (Dede, 2009) but also make VR a promising tool for early MCI detection. On the other hand, the Stroop test (Stroop, 1935), a well-established tool for assessing cognitive interference and response inhibition(Scarpina & Tagini, 2017), has been effectively combined with eye-tracking (ET) information to differentiate individuals with MCI from healthy controls (HC). Notably, MCI patients may exhibit distinctive eye movement patterns, *e.g.* prolonged fixation durations and impaired saccadic control, during Stroop tasks (Peltsch, Hemraj, Garcia, & Munoz, 2014). Moreover, ET signals, have shown promise as biomarkers for cognitive decline, with studies highlighting their ability to distinguish MCI patients from HC subjects (Opwonya, Doan, Kim, et al., 2022; Lagun, Manzanares, Zola, Buffalo, & Agichtein, 2011), as well as their potential for integration with machine learning methods to enhance detection accuracy (Song, Huang, Liu, et al., 2024). ET has also been demonstrated to effectively characterize everyday functional difficulties in MCI patients, *e.g.* impaired visual search and scene perception that are often overlooked by traditional assessments (Seligman & Giovannetti, 2015).

Building on these findings, we develop a VR-based cognitive assessment system that includes four tasks inspired by the Stroop test and integrates with the ET sensor. Our system leverages VR's ability to simulate real-world scenarios and ET's capacity to capture behavioral information of eyes, providing a comprehensive tool for assessing cognitive function. our method addresses the limitations of traditional methods and provides a more precise and user-friendly way for MCI assessment.

Despite the potential of ET signals in MCI detection, several challenges persist. ET signals are often non-stationary and exhibit significant temporal variability, making it diffi-

[1*] Corresponding authors 1.

[2*] Corresponding authors 2.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

cult to extract stable and discriminative features (Buettner, Scheuermann, Koot, Rössle, & Timm, 2018). Additionally, the high dimensionality and redundancy of ET features can lead to overfitting and increased computational complexity. For example, Lagun et al. (2011) demonstrated that traditional feature engineering methods may fail to capture subtle yet critical patterns in eye movements, while Song et al. (2024) highlighted the limitations of conventional machine learning models in handling the high-dimensional and noisy nature of ET signals. Furthermore, practical applications should consider lightweight deployment to ensure efficiency, particularly in resource-constrained environments.

To address these challenges, we propose the Kymatio Wavelet scattering (KWS) transform-based Temporal Attention (TA) with a Convolutional Neural Network (CNN), named KWS-TA-CNN, for MCI detection using ET signals. The main contributions are summarized as follows:

1. We developed a user-friendly VR cognitive assessment system with four Stroop-based tasks, integrated with a Tobii ET sensor to capture eye-behavioral data for MCI detection, resulting in an MCI dataset of 17 MCI and 21 HC subjects.
2. We propose a lightweight KWS-TA-CNN network: combining KWS for stable feature extraction, TA for enhanced discriminative representation, and 1D CNN for redundancy reduction, achieving improved MCI detection accuracy with computational efficiency.
3. The KWS-TA-CNN method demonstrates strong performance, with sample-level accuracies of 0.7629, 0.8304, and 0.7717, and subject-level accuracies of 0.8158, 0.9211, 0.8158, and 0.8421 across the four tasks in our MCI dataset.

Construction of MCI Dataset

VR-based cognitive assessment system

To facilitate the convenient and user-friendly assessment of cognitive states in the elderly, we developed a VR-based cognitive assessment system comprising both hardware and software components. The hardware setup includes an HP Reverb G2 Omnicept head-mounted display and two hand shanks, as shown in Figure 1 (a) and (b). The system features a resolution of 2160x2160 pixels per eye, a refresh rate of 90 Hz, and a field of view of 114°. Additionally, the system integrates a Tobii ET sensor that provides a sampling frequency of 120 Hz for ET signals. On the software side, we developed four tasks using the Unity3D platform, as depicted in Figure 1 (c1), (c2), (c3), and (c4). To enhance the sensitivity of cognitive assessment in detecting MCI, these tasks are categorized into consistent and inconsistent tasks based on the Stroop test (Stroop, 1935), with a pretest designed to help participants familiarize themselves with the tasks. The consistent tasks (Task A and Task B) involve simple text-matching and color-matching without interference, while the inconsistent tasks (Task C and Task D) introduce additional interference: Task C and Task D introduce

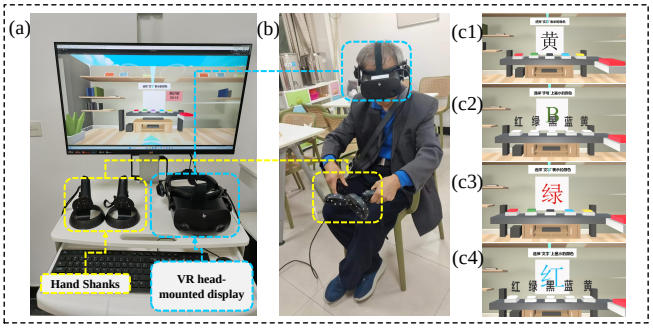

Figure 1: (a) HP Reverb G2 Omnicept head-mounted display and two hand shanks; (b) Participants; Four VR-based cognitive assessment tasks based-on the Stroop test: (c1) color-matching text without interference, (c2) text-matching color without interference, (c3) color-matching text with color interference, and (c4) text-matching color with text interference.

color interference and text interference to the baseline of Task A and Task B, respectively. Notably, for each task, ten cards are randomly presented to the participants, each card allowing 10 seconds for a response. Once the response is given or the time limit expires, the task automatically progresses to the next card or ends the task. This design increases the difficulty compared to the traditional color Stroop test, as participants are required to identify the correct answer and interact with the hand shanks in the system.

Experiment Procedure

Figure 2 illustrates the procedure for VR-based cognitive assessment, which consists of three phases for each participant: preparation, test, and self-report. The preparation phase (see Figure 2 (a)) involves participants completing an informed consent form, providing personal information, filling out the Mini-Mental State Examination (MMSE) scale (Folstein, Folstein, & McHugh, 1975), and calibrating the Tobii ET sensor while wearing the VR devices. This phase typically lasts around ten minutes. After successful calibration, participants proceed to the test phase, where they complete four tasks using the ET sensor for recording the ET signals (Figure 2 (b)), with the phase lasting approximately 13 minutes. A one-minute rest interval is provided after each task to minimize fatigue and enhance engagement, ensuring optimal performance throughout. In the self-report phase (Figure 2 (c)), participants complete a two-minute questionnaire assessing their subjective experiences, including VR usability, perceived task performance, and overall fatigue levels. The collected feedbacks can be used to exclude participants who report suboptimal experiences.

Subjects Recruitment

In this work, 45 participants were recruited to conduct the VR-based cognitive assessment to construct the MCI dataset from March to December 2024 at the Second Affiliated Hospital of Nanjing Medical University (Nanjing, China). In particular, none of the participants reported any ophthalmological

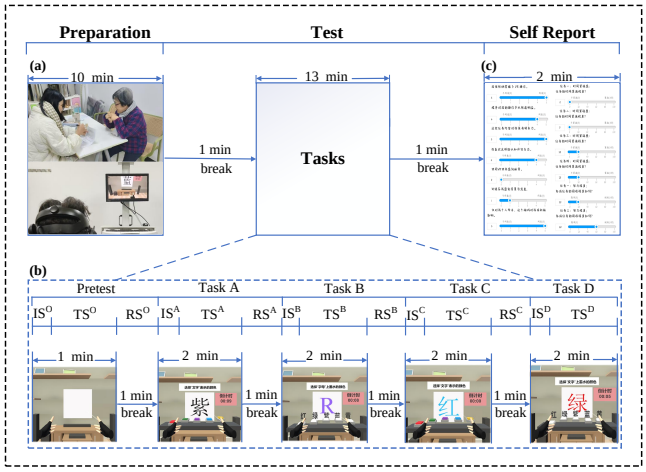

Figure 2: The procedure of VR-based cognitive assessment. (a) Preparation; (b) Test: a pretest and four tasks (IS: Introduction section, TS: Test section, RS: Rest section); (c) Self-report.

Table 1: The clinical information for MCI and HC groups in MCI dataset (mean±std), M: Male, F: Female. Edu. level: Educational level, where level=1, 2, and 3 indicate 1 to 5 years, 6 to 9 years, and high school diploma, repsectively.

Information	MCI (N=17)	HC (N=21)	Total (N=38)
Age	66.76±6.82	66.00±7.42	66.34±7.17
Gender	9(M)/8(F)	5(M)/16(F)	14(M)/24(F)
MMSE	26.12±1.78	27.90±1.51	27.10±1.86
Edu. level	2.65±0.59	1.95±0.72	2.29±0.75

conditions, aside from corrected vision. All subjects were informed and accepted the content and purpose of the experiment. Approval of all ethical, experimental procedures and protocols is granted by the Ethics Committee of Nanjing Medical University (No.2022784). After excluding 3 participants (failure in ET calibration) and 4 others (unfavorable subjective experiences as indicated by the questionnaire), the final MCI dataset comprised 38 participants. This cohort included 17 MCI patients and 21 HC subjects, aged between 55 and 82 years, with educational levels ranging from 1 to 3 (see Table 1 for details). To detect MCI patients and HC individuals, the study used a threshold τ for MMSE scores, adjusted based on educational level (Wang et al., 2011). The threshold is set as follows: $\tau = 23$ for Edu. level = 1, $\tau = 27$ for Edu. level = 2, and $\tau = 29$ for Edu. level = 3. Individuals scoring $< \tau$ were classified into the MCI group, while those scoring $\geq \tau$ were categorized into the HC group.

MCI Dataset

ET signals were continuously recorded throughout each task using the ET sensor in the VR device at a sampling rate of 120 Hz. The collected signals include 3D gaze vectors (normalized x, y, z, each $\in [-1, 1]$) and pupil positions (normalized x and y, each $\in [0, 1]$) for both eyes, along with pupil diameters (ranging from 2 to 8 mm), eye states (0: closed, 1: open), confidence levels (0: unreliable, 1: reliable), and missing data points (marked as -1) for each eye. To increase the richness and dimensionality of the ET signals, we also calculated the difference between the left-eye and right-eye signals by subtracting the right-eye values from the left-eye values, resulting in a total of 21-channel ET signals. Additionally, a pre-processing pipeline was applied to obtain clean signals, which included confidence-based filtering, linear interpolation to address blink-related missing signals, trimming of edge noise, and high-pass filtering (with a 1 Hz cut-off) to remove drift. To obtain enough samples, the clean 21-channel ET signals for each subject across each task are split into 2-second length (1-second overlapping) clean ET signals with a Hamming window operation. After pre-processing, the final number of samples for the four tasks in the MCI dataset is provided in Table 2.

Table 2: The number of samples for MCI and HC groups in MCI dataset.

Task	MCI (N=17)	HC (N=21)	Total (N=38)
A	996	1151	2147
B	906	1152	2058
C	750	993	1743
D	972	1167	2139
Total	3624	4463	8087

Methodology

As illustrated in Figure 3, for each Task i (i = A, B, C, D), our proposed KWS-TA-CNN network for MCI detection using ET signals in our constructed MCI dataset consists of three main components: (a) KWS is employed to extract the scattering feature matrix $\boldsymbol{K}^i$ from the clean signals $\boldsymbol{P}^i$, (b) TA dynamically weighs the important time steps, generating time-enhanced feature representation $\boldsymbol{T}^i$, and (c) CNN is used to capture local temporal patterns from $\boldsymbol{T}^i$, resulting in refined feature representation $\boldsymbol{O}^i$. Finally, $\boldsymbol{O}^i$ is fed into a fully connected layer for MCI or HC classification.

Kymatio Wavelet scattering transform

Inspired by the use of wavelet scattering feature for major depressive disorder recognition using multi-channel electroencephalogram signals (Zhang et al., 2025), which can effectively address the non-stationarity of signals, we adopt this approach in our method. However, it can be computationally expensive with high memory resources. To mitigate this, we utilize a high-performance implementation of KWS (Andreux et al., 2020) to compute the scattering coefficients efficiently.

The scattering transform (Bruna & Mallat, 2013) computes wavelet scattering coefficients through multiple layers, forming a hierarchical tree structure. Consider a set of 2-second clean ET signals $\boldsymbol{P}^i \in \mathbb{R}^{C \times L}$, where $C = 21$ is the number of channels, $L = 240$ is the number of sampling points. For

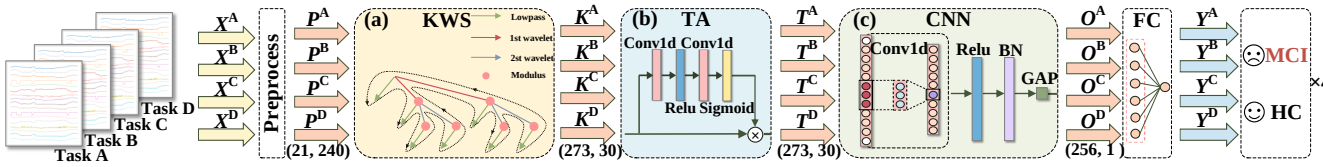

Figure 3: Overview of our proposed KWS-TA-CNN network for MCI detection using ET signals across four tasks. (a) Kymatio Wavelet scattering transform (KWS), (b) Time Attention (TA), (c) 1D Convolutional Neural Network (CNN).

each $\boldsymbol{P}^i$, the j-th channel signal is denoted as $\boldsymbol{P}^i_j \in \mathbb{R}^L$, where $j \in \{1,2,\ldots,C\}$. The 1D wavelet scattering transform with two-order wavelet scattering feature matrix K^i_j is defined as:

$$K^i_j = [K^i_{j(0)}; K^i_{j(1)}; K^i_{j(2)}] \tag{1}$$

The terms $K^i_{j(0)}$, $K^i_{j(1)}$, and $K^i_{j(2)}$ represent the 0th, 1st, and 2nd order scattering coefficients, respectively. $K^i_{j(0)}$ captures the low-frequency components of the signal P^i_j is computed as:

$$K^i_{j(0)} = P^i_j * \varphi_J \tag{2}$$

where φ_J is a Gaussian low-pass filter with scale parameter J and $*$ denotes convolution operation. $K^i_{j(1)}$ captures features at a specific frequency scale λ and is given as:

$$K^i_{j(1)} = |P^i_j * \psi_\lambda| * \varphi_J \tag{3}$$

where ψ_λ is the Morlet wavelet with frequency λ and $|\cdot|$ denotes the modulus operator. $K^i_{j(2)}$ is obtained to capture features across two different frequency scales λ and μ:

$$K^i_{j(2)} = |P^i_j * \psi_\lambda| * \psi_\mu| * \varphi_J \tag{4}$$

Specially, the convolution operations are efficiently computed using the Fourier Transform Convolution Theorem: $\boldsymbol{y} = \boldsymbol{x} * \boldsymbol{h} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x})\mathcal{F}(\boldsymbol{h}))$, where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote Fast Fourier transform (FFT) and inverse FFT, respectively. $\boldsymbol{x}$ and $\boldsymbol{h}$ represent the input signal and the filter, respectively.

Considering the signal length and the rapid changes in the ET signals, we set $J = 3$. The wavelet scattering transform of Eq. (1) outputs: $\boldsymbol{K}^i_{j(0)} \in \mathbb{R}^{N_0 \times T}$ with $N_0 = 1$ and $T = 30$, $\boldsymbol{K}^i_{j(1)} \in \mathbb{R}^{N_1 \times T}$ with $N_1 = 8$ and $\boldsymbol{K}^i_{j(2)} \in \mathbb{R}^{N_2 \times T}$ with $N_2 = 4$, resulting $\boldsymbol{K}^i_j \in \mathbb{R}^{N \times T}$ (where $N = N_0 + N_1 + N_2$). For each task i, $\boldsymbol{K}^i \in \mathbb{R}^{CN \times T}$ is constructed by stacking K^i_j of all channels: $\boldsymbol{K}^i = [\boldsymbol{K}^i_1; \boldsymbol{K}^i_2; \ldots; \boldsymbol{K}^i_C] \in \mathbb{R}^{CN \times T}$.

As illustrated in Figure 3 (a), KWS employs a depth-first traversal of the scattering transform tree, recursively processing each branch from the root to the deepest layer. This strategy reduces memory usage by avoiding the need to store all intermediate coefficients simultaneously, making it efficient for GPU execution.

Time Attention

As illustrated in Figure 3 (b), we proposed the TA to enhance the model's focus on critical time steps by applying lightweight operations: a 1×1 convolution for channel compression and another 1×1 convolution with a Sigmoid activation function for time step attention. This approach is inspired by Squeeze-and-Excitation Networks (SENet) (Hu, Shen, & Sun, 2018). The first 1×1 convolution operation compresses the channels of the input matrix $\boldsymbol{K}^i$ to obtain a compressed feature matrix $\boldsymbol{K}^{i'} \in \mathbb{R}^{CN/r \times T}$:

$$\boldsymbol{K}^{i'} = \boldsymbol{W}_1 \cdot \boldsymbol{K}^i + \boldsymbol{b}_1 \tag{5}$$

where r is the channel reduction ratio, $\boldsymbol{W}_1 \in \mathbb{R}^{CN/r \times CN}$ is the learnable weight matrix of the 1x1 convolution, and $\boldsymbol{b}_1 \in \mathbb{R}^{CN/r}$ is the bias term. Next, the attention weight matrix $\boldsymbol{A}^i \in \mathbb{R}^{1 \times T}$ is obtained through another 1×1 convolution operation:

$$\boldsymbol{A}^i = \sigma(\boldsymbol{W}_2 \cdot \boldsymbol{K}^{i'} + \boldsymbol{b}_2) \tag{6}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function, $\boldsymbol{W}_2 \in \mathbb{R}^{1 \times CN/r}$ is the learnable weight matrix of the second 1x1 convolution and $\boldsymbol{b}_2 \in \mathbb{R}$ is the bias term. Finally, the time-enhanced feature matrix $\boldsymbol{T}^i \in \mathbb{R}^{CN \times T}$ is obtained by:

$$\boldsymbol{T}^i = \boldsymbol{K}^i \odot \boldsymbol{A}^i \tag{7}$$

where $\odot$ denotes element-wise multiplication.

1D Convolutional Neural Network

To capture discriminative local temporal patterns and reduce feature redundancy, a 1D Convolutional Neural Network (CNN) is employed (LeCun, Bengio, & Hinton, 2015).

As shown in Figure 3 (c), the CNN includes the subsequent refinement steps (*e.g.* ReLU, normalization, dropout, and global average pooling), which can enhance the model's ability to learn discriminative feature vector $\boldsymbol{O}^i \in \mathbb{R}^H$, where H is the number of output channels (H = 256). Finally, $\boldsymbol{O}^i$ is fed into a fully connected layer for classification.

Experiment Results

The performance of KWS-TA-CNN network

To evaluate the performance of multi-channel ET signal-based MCI detection across the four VR-based tasks using our proposed KWS-TA-CNN method, we experiment with the leave-one-subject-out (LOSO) cross-validation strategy for each task. In this strategy, one subject from the MCI dataset is used as the test set, while the remaining subjects are utilized for training the model, repeating the process until every subject has been tested. Our experiment took place on a workstation with eight

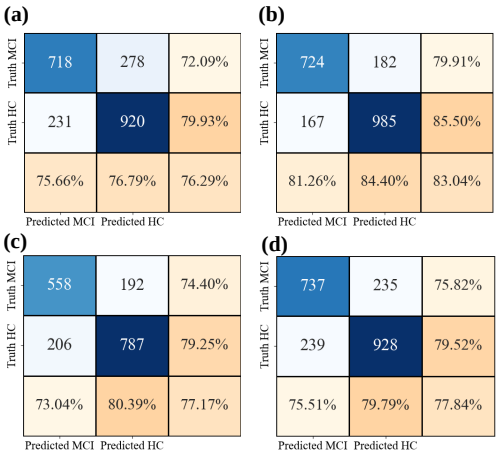

Figure 4: The sample-level confusion matrices for (a) Task A, (b) Task B, (c) Task C, and (d) Task D, respectively.

NVIDIA GeForce RTX 3090 GPUs. The Adam optimizer was used, with the learning rate set to 3e-6 and a weight decay of 1e-2. The number of epochs is uniformly set to 300 across all tasks, ensuring consistency in the training process for the model. We compute the accuracy (ACC), recall (REC), and precision (PRE) from the sample-level confusion matrix over all subjects, which aggregate counts across all samples. Additionally, the area under the curve (AUC) is calculated by aggregating the true labels and predicted probabilities from all cross-validation subjects.

Figure 4 presents the LOSO classification results at the sample level, where our proposed KWS-TA-CNN network achieves the best performance in Task B among the four tasks with ACC, REC, PRE of 0.8304, 0.7991, 0.8126, respectively. This is likely due to Task A being relatively easier for the subjects, while Tasks C and D posed more challenges.

For subject recognition in the LOSO experiment, the accuracies of each subject in MCI and HC groups across the four tasks are shown in Figure 5 (a-d). Most HC subjects achieve an accuracy of over 0.7 across all tasks (*e.g.*, 16 out of 21 for Task A). To simplify, we set a threshold accuracy of 0.5 for determining a subject's status, with subjects being classified based on whether their sample accuracy exceeded this threshold. The subject-level confusion matrices for the four tasks are shown in Figure 5 (e-h). The KWS-TA-CNN network demonstrates the best performance, achieving an accuracy of 0.9211 (see Figure 5 (f)). As also shown in Figure 5 (b), the only misclassifications are subject 8 and subject 16 from the MCI group, and subject 26 from the HC group, who are incorrectly identified as HC, HC, and MCI, respectively.

Ablation Study

Significance of Kymatio Wavelet scattering transform To validate the effectiveness of the KWS module, we compared our KWS-TA-CNN network with the TA-CNN network, which does not include the KWS module. The comparison results across the four tasks are given in Table 3. The results

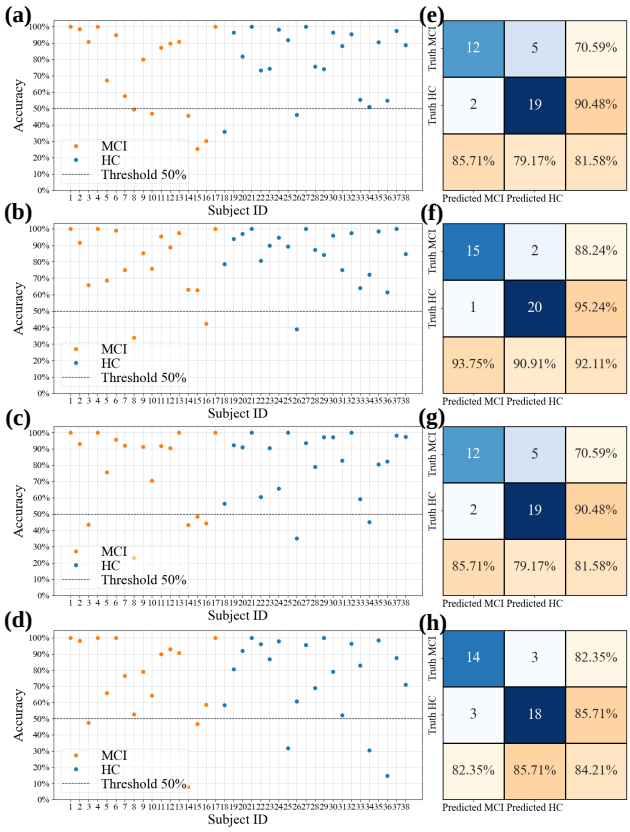

Figure 5: The accuracies for each subject in (a) Task A, (b) Task B, (c) Task C, and (d) Task D, respectively. The subject-level confusion matrices for (e) Task A, (f) Task B, (g) Task C, and (h) Task D, respectively.

reveal that including the KWS module significantly improves performance across all tasks. For instance, compared to the TA-CNN, the ACC and AUC of the KWS-TA-CNN network in Task B improves from 0.6088 to 0.8304 and from 0.6597 to 0.9082, respectively. Similarly, the ACC and AUC are improved by 0.2408 and 0.3665 for KWS-TA-CNN than ACC and AUC of 0.5376 and 0.4902 for TA-CNN in Task D, respectively. These results highlight the crucial role of KWS module in extracting time-robust feature, enhancing classification performance.

Significance of Time Attention To validate the effectiveness of the TA module, we also compared the KWS-TA-CNN network with the KWS-CNN network, which lacks the TA module. The comparison results across the four tasks are presented in Table 3. The results show that incorporating the TA module leads to task-dependent performance improvements. For instance, in Tasks B and C, the ACC improved from 0.8178 and 0.7705 for KWS-CNN to 0.8304 and 0.7717 for KWS-TA-CNN, while the AUC increased from 0.8975 and 0.8318 for KWS-CNN to 0.9082 and 0.8407 for KWS-TA-CNN, respectively. However, in Task D, the benefits of the

Table 3: The results of ablation study on KWS and TA modules across four tasks.

Task	Method	ACC	REC	PRE	AUC
A	TA-CNN	0.6075	0.5898	0.5753	0.6391
	KWS-CNN	**0.7634**	0.7068	**0.7652**	**0.8264**
	KWS-TA-CNN	0.7629	**0.7209**	0.7566	0.8214
B	TA-CNN	0.6088	0.6115	0.5501	0.6597
	KWS-CNN	0.8178	0.7826	0.7993	0.8975
	KWS-TA-CNN	**0.8304**	**0.7991**	**0.8126**	**0.9082**
C	TA-CNN	0.5846	0.5947	0.5150	0.5763
	KWS-CNN	0.7705	**0.7453**	0.7279	0.8318
	KWS-TA-CNN	**0.7717**	0.7440	**0.7304**	**0.8407**
D	TA-CNN	0.5376	0.5741	0.4925	0.4902
	KWS-CNN	**0.7873**	**0.7767**	**0.7603**	**0.8599**
	KWS-TA-CNN	0.7784	0.7582	0.7551	0.8567

TA module were less pronounced. These findings highlight the essential role of the TA module in dynamically adjusting temporal weights.

Table 4: The results of comparison study on different attention mechanisms across four tasks.

Task	Method	ACC	REC	PRE	AUC
A	KWS-SA-CNN	0.7382	0.6998	0.7260	0.8064
	KWS-CA-CNN	**0.7648**	0.7139	**0.7637**	**0.8244**
	KWS-TA-CNN	0.7629	**0.7209**	0.7566	0.8214
B	KWS-SA-CNN	0.7964	0.7627	0.7721	0.8772
	KWS-CA-CNN	0.8197	0.7870	0.8002	0.8921
	KWS-TA-CNN	**0.8304**	**0.7991**	**0.8126**	**0.9082**
C	KWS-SA-CNN	0.7602	0.7307	0.7173	0.8298
	KWS-CA-CNN	0.7653	**0.7520**	0.7166	0.8177
	KWS-TA-CNN	**0.7717**	0.7440	**0.7304**	**0.8407**
D	KWS-SA-CNN	0.7499	0.7078	0.7327	0.8369
	KWS-CA-CNN	0.7747	**0.7634**	0.7465	0.8439
	KWS-TA-CNN	**0.7784**	0.7582	**0.7551**	**0.8567**

Comparison study

Comparison of Attention Mechanisms To demonstrate the discrimination ability of the TA module, we compare our proposed KWS-TA-CNN network with two other attention mechanisms: 1) KWS-SA-CNN, which replaces the TA module with Self Attention (SA) module (Vaswani et al., 2017), and 2) KWS-CA-CNN, which incorporates Channel Attention (CA) based on SENet (Hu et al., 2018) adapted for 1D input. The ACC, REC, PRE, and AUC scores across all subjects for the three networks in the four tasks are provided in Table 4. The results show that the KWS-TA-CNN network outperforms the other two models in most tasks (Tasks B, C, and D) in terms of ACC, PRE, and AUC. For example, the ACCs of KWS-SA-CNN, KWS-CA-CNN, and KWS-TA-CNN are equal to 0.7964, 0.8197, and 0.8304 for Task B, 0.7602, 0.7653, and 0.7717 for Task C, and 0.7499, 0.7747, and 0.7784 for Task D, respectively.

Comparison of temporal classification networks To demonstrate the discrimination ability of CNN in extracting local features within the proposed network, we compare our KWS-TA-CNN network with three other temporal classification variants: (1) KWS-TA-LSTM, which replaces the CNN with a Long Short-Term Memory (LSTM) network (Gers & Schmidhuber, 2001); (2) KWS-TA-GRU, which uses a Gated Recurrent Unit (GRU) network (Chung, Gulcehre, Cho, et al., 2014) instead of the CNN; and (3) KWS-TA-TCN, which employs a Temporal Convolutional Network (TCN) (Bai, Kolter, & Koltun, 2018) as the classification backbone. The performance results of the four networks are shown in Table 5. As seen in the table, KWS-TA-CNN consistently outperforms the three variants across all tasks, except for the PRE in Task A, highlighting the superior ability of CNN to capture local patterns. Remarkably, KWS-TA-CNN achieves an accuracy of 0.8304 for Task B, significantly surpassing KWS-TA-LSTM (0.7517), KWS-TA-GRU (0.8042), and KWS-TA-TCN (0.7201).

Table 5: The results of comparison study on different temporal classification networks across four tasks.

Task	Method	ACC	REC	PRE	AUC
A	KWS-TA-LSTM	0.7536	0.6888	**0.7580**	0.7904
	KWS-TA-GRU	0.7313	0.6677	0.7300	0.7178
	KWS-TA-TCN	0.6232	0.5100	0.6128	0.6891
	KWS-TA-CNN	**0.7629**	**0.7209**	0.7566	**0.8214**
B	KWS-TA-LSTM	0.7517	0.6932	0.7294	0.8112
	KWS-TA-GRU	0.8042	0.7517	0.7928	0.8200
	KWS-TA-TCN	0.7201	0.6700	0.6867	0.7091
	KWS-TA-CNN	**0.8304**	**0.7991**	**0.8126**	**0.9082**
C	KWS-TA-LSTM	0.6856	0.6000	0.6447	0.7218
	KWS-TA-GRU	0.7435	0.6627	0.7192	0.6989
	KWS-TA-TCN	0.6902	0.5760	0.6606	0.6092
	KWS-TA-CNN	**0.7717**	**0.7440**	**0.7304**	**0.8407**
D	KWS-TA-LSTM	0.7695	0.7459	0.7467	0.8402
	KWS-TA-GRU	0.7532	0.7160	0.7342	0.7801
	KWS-TA-TCN	0.7204	0.7243	0.6809	0.8246
	KWS-TA-CNN	**0.7784**	**0.7582**	**0.7551**	**0.8567**

Conclusion

In this study, we propose a novel lightweight network, KWS-TA-CNN, for detecting MCI using ET signals in VR-based cognitive assessment tasks. The network integrates the KWS, TA, and CNN components, which collectively address the challenges of non-stationarity, temporal variability, and redundancy in MCI classification using the ET signals. Experimental results using the LOSO cross-validation strategy demonstrate strong performance, with high classification accuracies both at the sample level and at the subject level across four tasks. These results highlight the model's potential for MCI detection and its clinical applicability.

However, this study has some limitations. The relatively small number of subjects in the dataset, may affect the model's generalization and performance. Additionally, there is a need to explore methods for handling noisy or incomplete data to further improve the model's performance.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2023YFC3603600) and the National Natural Science Foundation of China under Grants (62001240, 31400842).

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... others (2013). The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Focus*, *11*(1), 96–106.

Andreux, M., Angles, T., Exarchakis, G., Leonarduzzi, R., Rochette, G., Thiry, L., ... others (2020). Kymatio: Scattering transforms in python. *Journal of Machine Learning Research*, *21*(60), 1–6.

Arbizu, J., Festari, C., Altomare, D., et al. (2018). Clinical utility of FDG-PET for the clinical diagnosis in MCI. *European Journal of Nuclear Medicine and Molecular Imaging*, *45*, 1497–1508.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1872–1886.

Buettner, R., Scheuermann, I. F., Koot, C., Rössle, M., & Timm, I. J. (2018). Stationarity of a user's pupil size signal as a precondition of pupillary-based mental workload evaluation. In *Information systems and neuroscience: Gmunden retreat on neurois 2017* (pp. 195–200).

Chung, J., Gulcehre, C., Cho, K., et al. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66–69.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189–198.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., ... others (2006). Mild cognitive impairment. *The Lancet*, *367*(9518), 1262–1270.

Gers, F. A., & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, *12*(6), 1333–1340.

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).

Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., & Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of Neuroscience Methods*, *201*(1), 196–203.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Mirzaei, G., & Adeli, H. (2022). Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia. *Biomedical Signal Processing and Control*, *72*, 103293.

Mofrad, S. A., & et al. (2021). A predictive framework based on brain volume trajectories enabling early detection of alzheimer's disease. *Computerized Medical Imaging and Graphics*, *90*, 101910.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., ... Chertkow, H. (2005). The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699.

Opwonya, J., Doan, D. N. T., Kim, S. G., et al. (2022). Saccadic eye movement in mild cognitive impairment and alzheimer's disease: A systematic review and meta-analysis. *Neuropsychological Review*, *32*, 193–227.

Peltsch, A., Hemraj, A., Garcia, A., & Munoz, D. P. (2014). Saccade deficits in amnestic mild cognitive impairment resemble mild Alzheimer's disease. *The European Journal of Neuroscience*, *39*(11), 2000–2013.

Ritchie, C., Smailagic, N., Noel-Storr, A. H., Takwoingi, Y., Flicker, L., Mason, S. E., & McShane, R. (2014). Plasma and cerebrospinal fluid amyloid beta for the diagnosis of alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*(6).

Scarpina, F., & Tagini, S. (2017). The stroop color and word test. *Frontiers in Psychology*, *8*, 557.

Scheltens, P., Blennow, K., Breteler, M. M., de Strooper, B., Frisoni, G. B., Salloway, S., & Van der Flier, W. M. (2016). Alzheimer's disease. *The Lancet*, *388*(10043), 505–517.

Seligman, S. C., & Giovannetti, T. (2015). The potential utility of eye movements in the detection and characterization of everyday functional difficulties in mild cognitive impairment. *Neuropsychology Review*, *25*(2), 199–215.

Siuly, S., Alçin, F., Kabir, E., Şengür, A., Wang, H., Zhang, Y., & Whittaker, F. (2020). A new framework for automatic detection of patients with mild cognitive impairment using resting-state eeg signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(9), 1966-1976.

Song, J., Huang, H., Liu, J., et al. (2024). Diagnostic potential of eye movements in Alzheimer's disease via a multiclass machine learning model. *Cognitive Computation*, *16*, 3364–3378.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

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

Wang, B., Guo, Q.-H., Chen, M.-R., Zhao, Q.-H., Zhou, Y.,

& Hong, Z. (2011). The clinical characteristics of 2,789 consecutive patients in a memory clinic in china. *Journal of Clinical Neuroscience*, *18*(11), 1473–1477.

Zhang, F., Yang, C., You, L., Wang, X., Yuan, Y., Le Bouquin Jeannès, R., … Xiang, W. (2025). Ws-bilstm-ma: Wavelet scattering-based bilstm with mixed attention block for mdd recognition using multichannel eeg signals. *IEEE Transactions on Instrumentation and Measurement*, *74*, 1-13.