An Effective Entropy-Assisted Mind-Wandering Detection System Using EEG Signals of MM-SART Database

Yi-Ta Chen, *Student Member, IEEE*, Hsing-Hao Lee, Ching-Yen Shih, Zih-Ling Chen, Win-Ken Beh, *Student Member, IEEE*, Su-Ling Yeh, and An-Yeu Wu, *Fellow, IEEE*

***Abstract*—Mind-wandering (MW), which is usually defined as a lapse of attention has negative effects on our daily life. Therefore, detecting when MW occurs can prevent us from those negative outcomes resulting from MW. In this work, we first collected a multi-modal Sustained Attention to Response Task (MM-SART) database for MW detection. Eighty-two participants' data were collected in our dataset. For each participant, we collected measures of 32-channels electroencephalogram (EEG) signals, photoplethysmography (PPG) signals, galvanic skin response (GSR) signals, eye tracker signals, and several questionnaires for detailed analyses. Then, we propose an effective MW detection system based on the collected EEG signals. To explore the non-linear characteristics of the EEG signals, we utilize entropy-based features. The experimental results show that we can reach 0.712 AUC score by using the random forest (RF) classifier with the leave-one-subject-out cross-validation. Moreover, to lower the overall computational complexity of the MW detection system, we propose correlation importance feature elimination (CIFE) along with AUC-based channel selection. By using two most significant EEG channels, we can reduce the training time of the classifier by 44.16%. By applying CIFE on the feature set, we can further improve the AUC score to 0.725 but with only 14.6% of the selection time compared with the recursive feature elimination (RFE). Finally, we can apply the current work to educational scenarios nowadays, especially in remote learning systems.**

Manuscript received 10 June 2021; revised 20 April 2022 and 24 June 2022; accepted 26 June 2022. Date of publication 29 June 2022; date of current version 9 August 2022. This research was supported by the Ministry of Science and Technology of Taiwan (MOST) under Grants MOST-107-2410-H-002-129-MY3, MOST-107-2622-8-002-009-TA, MOST-108/109-2622-8-002-012-TA, and MOST-110-2410-H-002-130-MY3.

Yi-Ta Chen and Win-Ken Beh are with the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei City 10617, Taiwan (e-mail: edan@access.ee.ntu.edu.tw; kane@access.ee.ntu.edu.tw).

Hsing-Hao Lee is with the Department of Psychology, National Taiwan University, Taipei City 10617, Taiwan (e-mail: hsinghaolee@gmail.com).

Ching-Yen Shih is with the Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: cys@access.ee.ntu.edu.tw).

Zih-Ling Chen is with the Graduate Institute of Brain and Mind Sciences, National Taiwan University College of Medicine, Taipei City 10617, Taiwan (e-mail: annie0191@hotmail.com).

Su-Ling Yeh is with the Department of Psychology, National Taiwan University, Taipei City 10617, Taiwan (e-mail: suling@ntu.edu.tw).

An-Yeu Wu is with the Graduate Institute of Electronics Engineering, National Taipei University, Taipei City 10617, Taiwan (e-mail: andywu@ntu.edu.tw).

Digital Object Identifier 10.1109/JBHI.2022.3187346

***Index Terms*—Correlation importance feature selection (CIFE), EEG, entropy features, mind-wandering, sustained attention to response task (SART).**

I. Introduction

MIND-WANDERING (MW) is the experience of thoughts not remaining on a single topic, particularly when people are engaged in an attention-demanding task [1]. It has been estimated that the occurrence of MW is between 20%–40% in our daily life [2]. If MW occurs during driving, it might put people in danger [3]. Also, MW will degrade the efficiency of learning if it appears when people are studying [4], especially during the COVID-19 pandemic where on-line learning becomes a necessity. In addition, when people mind-wander, they are unhappier than when they focus on the task at hand [5]. These negative emotions worsen the mental health of modern people immersed in the negative information from the COVID-19 pandemic [6]. Therefore, analyzing and detecting MW with efficient mechanism is of great interest in recent days.

There were several experiments designed to induce MW. In [7], they used the Sustained Attention to Response Task (SART) to analyze the everyday attentional failures and action slips between brain-injured patients and normal participants. In [8], the authors designed the experiment by collecting eye gaze data and galvanic skin response of an Affectiva Q sensor while participants were reading texts. The authors of [9] designed a simulated driving experiment to detect MW by EEG signals and labeled the data by auditory probes. The aforementioned studies have shown that MW can be induced by proper experimental designs, and states of attention can be captured by certain physiological measurements.

Since MW is a mental state defined as the lapse of attention, it is highly correlated with the internal activities of our brain. As a result, researchers often study the relationship between EEG signals and MW. For example, event-related potentials (ERP) were used to explore the effect of MW on processing relevant or irrelevant events [10]. In [9], the authors observed increased power in the α band during MW periods. In a current study [11], the author collected data from 30 subjects using the SART experiment design. Moreover, they analyzed the collected EEG data using support vector machine with ERP and time-frequency features. They found out that alpha power is most predictive of

MW. Most of the studies applied band power and ERPs as the major features. However, note that the EEG signals have long-term memory effects and spatial cross-dependencies, which require non-linear methods to extract the non-linear information. The authors of [12] modeled this non-linearity with complex network models and reported meaningful results with less features. Moreover, EEG signals can be described and forecasted with a spatial-temporal fractal model [13] that relies on few parameters to estimate the unknown stimuli [14]. Most works illustrated and supported the importance of the non-linearity features of EEG. That also motivates us to exploit more non-linear features in the entropy domain.

In recent years, several non-linear features of EEG were proposed. For example, the authors employed the permutation entropy to perform complexity analysis of Alzheimer's disease [15]. Wavelet entropy features were applied in discriminating patients with attention deficit hyperactivity disorder (ADHD) from healthy controls, and a 96% accuracy was achieved in this task [16]. In [17], Fractal Dimension, as a type of complexity feature, was also treated as a biomarker for dementia. These related works have demonstrated that the complexity information of EEG can perform well in the classification tasks. However, none of them has adopted entropy in characterizing human attentional functions. Hence, it is desirable to extract entropy-based features to detect MW using EEG data.

Additionally, due to the multi-channel characteristics and large number of extracted features, the overall computational complexity of the system becomes extremely high. Therefore, several simplification methods were proposed to lower the total system complexity. In [18], the authors used features with p-values smaller than 0.05 in the emotion recognition task to reduce the total number of features. This not only reduces the computational complexity of the classifier, but also helps analyze the dominant features to support their study. The authors of [19] proposed multiscale permutation entropy (MPE) to reduce the complexity of original multiscale entropy (MSE) [20]. Those simplified techniques can lower the total complexity and further improve the overall efficiency of the system.

In order to build a MW detection system, several issues need to be considered:

1) *Lacking a comprehensive database for MW detection:* There are some desirable features to build a complete MW database: First, the generalization of the database is important. Second, multi-modality is required to evaluate the effectiveness of different physiological signals. Third, large number of subjects is critical to the confidence of detection performance. There are already several experiments designed to collect data for detecting MW [2], [8], [9], [21]. Nevertheless, they cannot cover all the desired features.
2) *Lacking the exploration of non-linear EEG features for MW detection:* Many studies have focused on the processing of EEG in MW detection. However, most of them applied basic EEG features on the time or frequency domain to reveal the differences between MW and non-MW. Due to the non-linearity of EEG, the effectiveness of non-linear features, such as entropy-based features, should be explored for detecting MW.
3) *High computation complexity of EEG-based MW detection system:* While dealing with EEG signals, the total number of features is relatively large due to the many channels of EEG. Moreover, the computational complexity of several entropy-based features is $O(n^2)$ [22]. Hence, if we directly extract and apply all EEG features to the MW detection system, the overall computational complexity will be a big burden. Hence, a efficient MW detection system should be proposed.

To tackle the above issues, we collect a new multi-modal Sustained Attention to Response Task (MM-SART) database, and design an effective infrastructure of detecting MW based on EEG signals. The main contributions of this paper are as follows:

1) *Establish the Multi-modal Sustained Attention to Response Task (MM-SART) open database:* To support our research on MW detection based on physiological signals, we design and collect a new Multi-Modal Sustained Attention to Response Task (MM-SART) database. Multi-modality physiological signals are collected in a controlled environment to exclude external factors. Up to now, we have collected data from 82 participants to support our analysis. To the best of our knowledge, the MM-SART database is the first database with the most modalities and the largest number of subjects for MW detection. Moreover, the MM-SART database will be open-sourced. For researchers who are interested in doing experiments for MW detection, they can easily access the MM-SART database [23] (URL: http://mmsart.ee.ntu.edu.tw/.)
2) *Apply entropy-based features to MW detection:* We propose to extract non-linear information of EEG based on entropy-based features. We analyze entropy-based features on time, frequency, and wavelet domain, respectively. From the experimental results, we show that the extracted entropy-based features are complementary to traditional linear features (e.g., statistical and band power features). By utilizing the new entropy-domain features, we can reach a 0.670 F1 score, 0.318 Kappa score, and 0.712 Area-under-Curve (AUC) score in the leave-one-subject-out cross-validation. Comparing with the performance of utilizing only the basic statistical features, which is 0.630 F1 score, 0.237 Kappa score, and 0.677 AUC, we can improve our performance by 0.04 F1 score, 0.081 Kappa score, and 0.035 AUC.
3) *Improve the computational efficiency of EEG-based MW detection system:* We propose to apply channel selection and feature selection to the EEG data. We firstly apply the AUC-based channel selection to reach the optimal point between the number of channels and the AUC score. Secondly, we propose a correlation importance feature elimination (CIFE) based on the Random Forest (RF) classifier to select the most significant features. By applying two most significant EEG channels to the MW detection system, we can reduce the training time of the classifier by 44.16% with only 0.016 degradation of the AUC score. By performing CIFE on the feature set, we can further improve the AUC score to 0.725 with only 14.6% of the selection time compared with the recursive feature elimination (RFE). From the top-14 features selected

from CIFE, there are 5 entropy-based features from different categories, such as the T7_MFDE, T7_WL-MFDE, FP2_MPE, FP2_WL-Ent, and FP2_cD7-WL-SpecEnt, which shows the effectiveness of our proposed entropy-based features.

The rest of the paper is organized as follows. In Section II, the MM-SART database is introduced. The proposed effective MW detection system based on EEG is presented in Section III. In Section IV, the experiment result of the effective MW detection system is shown. Channel selection and feature selection of the EEG-based MW detection system are introduced in Section V. Finally, the conclusions are drawn in Section VI.

II. Multi-Modal Sustained Attention to Response Task (MM-SART) Open Database

A. Background

To detect and analyze the process of MW, a proper experiment with a well-controlled environment should be proposed. Moreover, the source used to detect MW is important. MW is related to several neural processes [24], such as increased activities in the default mode network (DMN), suppressed activities within the anti-correlated (task-positive) network (ACN), as well as other changes in neuromodulations. Previous studies have utilized these brain connections to build up a model to detect states of attention [24]. However, using functional magnetic resonance imaging (fMRI) is less flexible and portable, which cannot be easily utilized in daily situations. Therefore, detecting MW by multi-modality physiological signals is more applicable to educational scenarios as the current study has done.

In our experiments, we adopt a modified version of the SART. We used a pseudo-random probe-based method to access the mental state of subjects. Moreover, we collected a multi-modality database to support various kinds of research. The website of the complete MM-SART database can be found in: http://mmsart.ee.ntu.edu.tw/.

B. Participants

82 participants were recruited in the current study. Five participants were excluded from the data analysis due to technical issues. Therefore, we ended up with 77 participants (age range: 20–33 years old, 40 females). All of the participants were right-handed and free from psychological and neurological disorders. They all had normal or corrected-to-normal vision. The experiment was approved by the Research Ethics Committee at National Taiwan University (NTU REC: 201812HM004) and executed with the compliance to the guidelines.

C. Apparatus

All stimuli were presented in a gray background on an ASUS 22” LED monitor with a spatial resolution of 1920 × 1080 pixels. EEG, eye tracker signals, and physiological signals were recorded by 3 systems, respectively. Stimuli were presented with E-prime (Psychology Software Tools, Pittsburgh, PA, USA), and triggers were also sent by E-prime and synchronized with a DB-25 connector.

EEG data were recorded with Neuroscan (El Paso, TX, USA) with 32-channel Quick-cap (AgCl electrodes). The recordings were originally referenced to the left mastoid (M1), and re-referenced to the average of the left and right mastoid (M2) offline. Vertical electrooculogram (V-EOG) was recorded from participants’ left eye with two electrodes (one placing on approximately 2 cm above the left eye, and the other was 2 cm below the left eye). Horizontal electrooculogram (H-EOG) was recorded with pairs of electrodes placing at 2 cm away from the left and right eye respectively. Before starting the experiment, the impedances of all electrodes were kept under to ensure the quality of data. EEG and EOG signals were amplified by the SynAmps using a 0.05–100 Hz bandpass and continuously sampled at 1000 Hz per channel for offline analysis.

All participants’ heart rate, skin conductance, skin temperature, and respiration data were recorded by the ProComp Infiniti (ProComp Infiniti of Thought Technology Ltd) at 2048 Hz and downsampled to 256 Hz while exporting the data. In addition, their eye movements data were recorded by Tobii Eye Glasses 2 (Tobii Technology, Danderyd, Sweden) with the sampling rate of 100 Hz.

D. Stimuli and Design

Participants were seated in a sound-attenuated room with their eyes approximately 80cm from the monitor. They were instructed to do the SART proposed in [25].

In this task, each block includes 25 trials and a probe at the end of the block. Participants were instructed to press number 9 on the number-pad with their right hand to initiate a block. Each block was embedded with 25 English letters (A-Y) in a pseudo-random order with one target letter (i.e., letter C, and the target probability was 4%), which appeared pseudo-randomly at one of the trials between the 6th and 15th trial in a block.

Participants were instructed to press number 8 with their right hand as soon as possible when they caught sight of a non-target letter but to withhold their response when they see the target letter “C”. Each letter was presented for 2000 ms or until the participant responded. The inter-trial interval (ITI) varied with the reaction time (RT) of participants so that each trial (including ITI) lasted for 2000 ms. For example, if the participant’s response time was 300 ms, then the ITI would be 1700 ms to equate the duration of each trial. There were 40 blocks in total.

E. Procedure

The procedure of the overall experiment is shown in Fig. 2(b). After signing the informed consent, participants were instructed to fill in the questionnaires. After that, they are equipped with bio-sensors from ProComp Infiniti to collect photoplethysmography (PPG) signals, galvanic skin response (GSR) signals, skin temperature on their left-hand fingers and the respiration data with the respiration sensor on their upper abdomen. We also used Neuroscan’s EEG cap to collect the EEG signals, and Tobii Eye Glasses 2 to collect the eye tracking signals. Total six kinds of signals are collected. Next, participants were recorded during a 3-minute closed-eye and open-eye resting-state. Then, participants were instructed to practice for 3 blocks to make sure

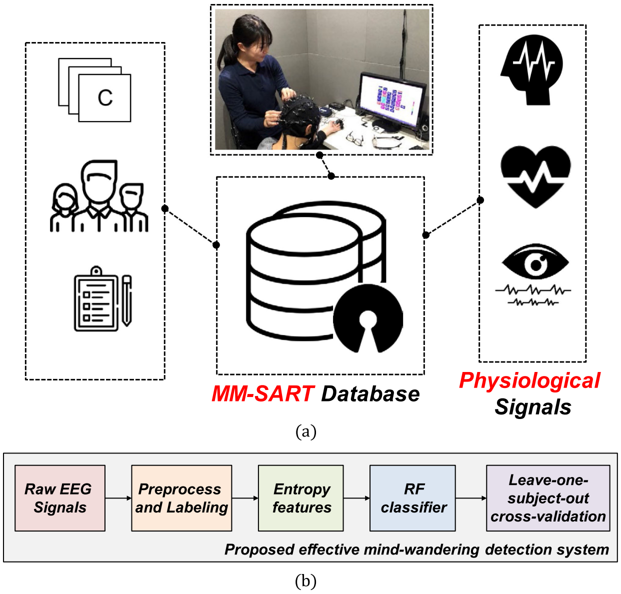

Fig. 1. Overview of (a) Proposed MM-SART database, and (b) Proposed Effective mind-wandering detection system based on EEG signals.

that they understood the SART. Before the formal experiment started, participants were asked to rate their sleepiness on a 4-point Likert scale (from 1: very alerted to 4: very sleepy) of their current state. Participants were told to do the task at their own pace and were allowed to rest at the end of each block. After the formal experiment ended, participants were asked to rate their state of sleepiness again on the 4-point Likert scale, followed by closed-eye and open-eye resting-state signals recording.

F. Self-Assessment of Participants

Self-assessment has been widely used in the research of mind-wandering [26]. In our experiment, a probe popped out and asked participants to classify the content of their thoughts with the question "What was in your mind just now?" with five options (1. Focusing on the task; 2. Thinking of the task performance; 3. Distracted by task-unrelated stimuli; 4. Thinking of things unrelated to the task; and 5. Nothing in particular) at the end of the block. After classifying their thought contents, another rating question asked participants to subjectively rate their state of focus from 1 (completely wandering) to 7 (very focused) at the moment before seeing the probe. Participants were told that they should respond truthfully and that there was no correct answer for the probe and the rating questions.

III. Proposed EEG-Based Mind-Wandering Detection System

In this work, we focus on designing a MW detection system based on the EEG signal. The use of multi-modality signals can be extended based on this initial study of the MM-SART database. From several related works, the processing flow of the EEG can be separated into few steps as shown in Fig. 3. The details of each building block will be described as follows.

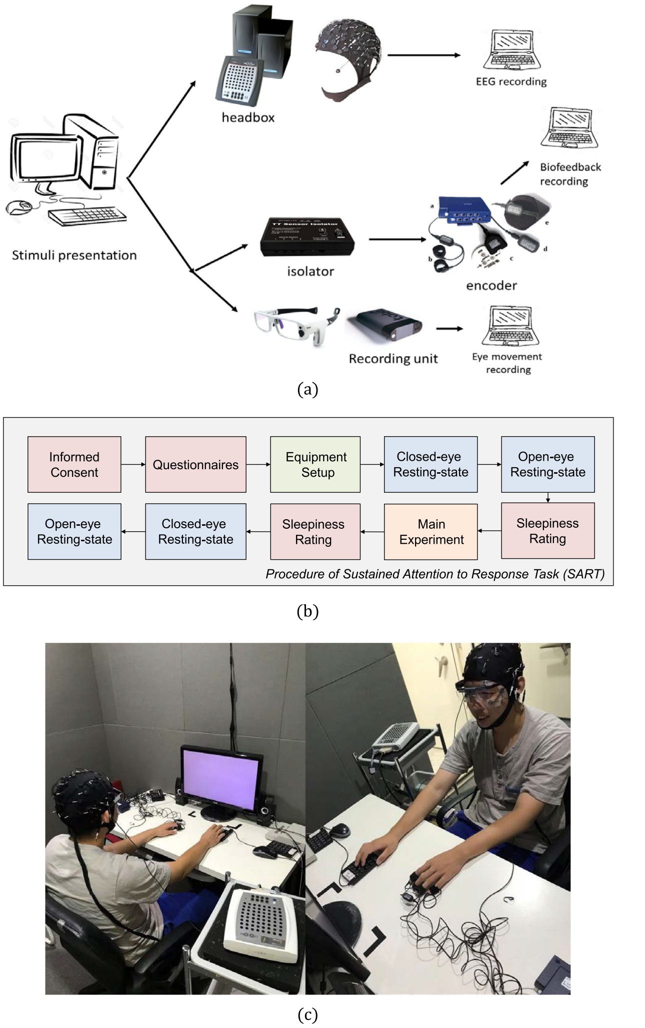

Fig. 2. (a) Data collection and synchronization among Neuroscan, Tobii Eye Glasses and ProComp Infiniti bio sensor system. (b) Procedure of the SART experiment. (c) The demonstration of a participant performing the task.

A. Preprocessing and Labeling

In the preprocessing step, we re-reference all EEG channels according to the average of the M1 and M2 channels. We use a bandpass filter from 0.1 Hz to 45.0 Hz to remove the baseline and high-frequency noise. Eye artifacts are noise when classifying EEG signals [9], and hence independent component analysis (ICA), such as FastICA algorithm [27], is usually applied to the EEG to remove eye artifacts. However, according to [8], eye blink information is useful for detecting MW. Therefore, we propose to process the EEG signals without ICA to utilize the information of eye movements in EEG signals.

Finally, we segment 10-s EEG signals before probes into epochs for each subject to predict the mental state. The labeling of each segment is based on the self-assessment score. In this paper, we consider the 7-point Likert scale as our labeling target

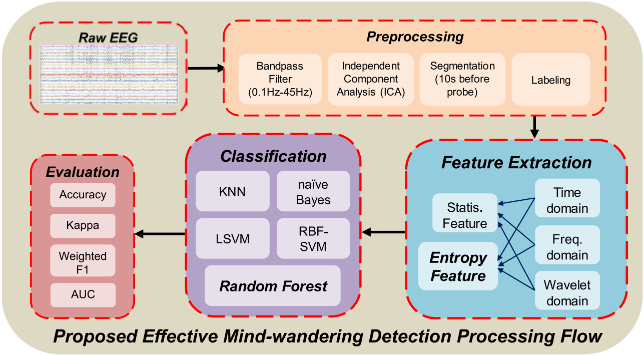

Fig. 3. The processing flow of EEG signals for the MM-SART database.

TABLE I
STATISTICAL FEATURES AND ENTROPY-BASED FEATURES IN THREE DIFFERENT DOMAINS

Feature sets (Feature number)	Time	Frequency	Wavelet
Statistical	Mean, Mean Power, 1^{st} diff, Hjorth complexity (5)	Power spectral density of θ, α, β, γ bands (4)	Mean power, Mean, Standard deviation, a ratio of absolute mean values of adjacent bands Mean power, Mean, Standard deviation, a ratio of absolute mean values of adjacent bands (20)
Entropy	MSE, MPE, MDE, MFDE (80)	Spectral entropy (5)	MPE, MDE, MFDE (310)

to evaluate the subjective MW state. Moreover, trials scoring 4 are removed because they cannot be categorized into either of the two states. After labeling each trial, we remove 21 subjects who always labeled themselves as MW or not MW.

B. Feature Extraction

We extract statistical and entropy-based features, which are summarized in Table I. As for the entropy-based features, we extract the multiscale entropy (MSE) [20], the multiscale permutation entropy (MPE) [19], the multiscale dispersion entropy (MDE) [28], and the multiscale fluctuation-based dispersion entropy (MFDE) [29] to capture the complexity of EEG signals in different scales in time and wavelet domains. These non-linear entropy-based features are introduced as follows. Note that the "multiscale" means coarse-graining process before the entropy calculation.

1) Multiscale Entropy (MSE): The extraction of multiscale entropy (MSE) [20] consists of two steps. The first step is the coarse-graining process. For a given time series $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$, the coarse-graining process will average the data points within a non-overlapping window of length τ, where τ is the scale factor. Each element of the coarse-grained series, $y_j^{(\tau)}$, is represented as:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i\ =(j-1)\ \tau+1}^{j\tau} x_i. \tag{1}$$

The second step of MSE extraction is the sample entropy calculation on the coarse-grained time series. This step aims to capture the probability of new pattern generated in the coarse-grained time series. The higher value the sample entropy is, the higher probability of new pattern generation. The sample entropy calculation step is defined as follows:

$$SampEn\ \left(y^{(\tau)}, m, \gamma\right) = -\ln \frac{n^{(m+1)}}{n^{(m)}}, \tag{2}$$

where $y^{(\tau)}$ represents the coarse-grained time series, $n^{(m)}$ is the number of matched patterns with dimension m in $y^{(\tau)}$, and γ is the maximum matching tolerance, which is set to 0.1 times standard deviation of the time series in our experiment.

2) Permutation Entropy (PE): Permutation entropy (PE) [19] is based on the counting of ordinal patterns that describe the up-and-down in the signals. The permutation pattern is denoted as a motif that uses relative order to indicate different kinds of amplitude variation of the signals. Specifically, with the pattern of dimension m, there are $m!$ distinct permutation patterns $\{\pi_1, \pi_2, \ldots, \pi_{m!}\}$ in the signal $\boldsymbol{x}$ of length N. The probability of each pattern is defined as:

$$p\left(\pi_j\right) = \frac{\#\{i|0 < i < i - m, (x_{i+1}, \ldots, x_{i+m})\ has\ type\ \pi_j\}}{N - m + 1}, \tag{3}$$

and the permutation entropy value is calculated based on the Shannon's definition of entropy as:

$$PermEn\ (x, m) = -\sum_{j=1}^{m!} p\left(\pi_j\right) \ln p\left(\pi_j\right). \tag{4}$$

3) Dispersion Entropy (DE): Dispersion entropy (DE) [28] can detect noise bandwidth, simultaneous frequency, and amplitude change. The dispersion entropy calculation consists of four steps. In the first step, the time series $\boldsymbol{x}\ = \{x_1, x_2, \ldots, x_N\}$ is mapped to c classes. By employing the normal cumulative distribute function (NCDF):

$$y_j = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_j} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt, \tag{5}$$

we can map $\boldsymbol{x}$ into $\boldsymbol{y}\ = \{y_1, y_2, \ldots, y_N\}$. Next, we use a linear algorithm to map each y_j to $z_j^{(c)}$ according to

$$z_j^{(c)} = round\left(c \cdot y_j + 0.5\right), \tag{6}$$

where $z_j^{(c)}$ is an integer from 1 to c.

In the second step, the embedding vector $z_i^{(m,c)}$ with embedding dimension m and time delay d is created according to

$$\boldsymbol{z}_i^{(m,c)} = \left\{z_i^{(c)}, z_{i+d}^{(c)}, \ldots, z_{i+(m-1)d}^{(c)}\right\},\ i = 1, 2, \ldots, N - (m-1)\, d. \tag{7}$$

The time series $\boldsymbol{z}_i^{(m,c)}$ is mapped to the dispersion pattern $\pi_{v_0 v_1 \ldots v_{m-1}}$, where $v_0 = z_i^{(c)}$, $v_1 = z_{i+d}^{(c)}$, . . . , $v_{m-1} = z_{i+(m-1)d}^{(c)}$. The number of possible dispersion patterns is equal to c^m.

In the third step, the probability of each pattern is calculated as:

$$p\left(\pi_{v_0 v_1 \ldots v_{m-1}}\right) = \frac{\#\left\{i | i < N - (m-1)\,d, \boldsymbol{z}_i^{(m,c)}\ has\ type\ \pi_{v_0 v_1 \ldots v_{m-1}}\right\}}{N - (m-1)\,d}. \tag{8}$$

In the last step, the dispersion entropy with embedding dimension m, time delay d, and the number of classes c is defined as follows:

$$DispEn\ (x, m, c, d) = -\sum_{\pi=1}^{c^m} p\left(\pi_{v_0 v_1 \ldots v_{m-1}}\right) \ln p\left(\pi_{v_0 v_1 \ldots v_{m-1}}\right). \tag{9}$$

4) Fluctuation-Based Dispersion Entropy (FDE): Fluctuation-based dispersion entropy (FDE) [29] is more stable than dispersion entropy over the irrelevant local trend. The main difference between FDE and DE is the second step mentioned in the dispersion entropy. The FDE considers the differences between adjacent elements of dispersion patterns. Thus, each element in the fluctuation-based dispersion pattern changes from $-c+1$ to $c-1$, and there are $(2c-1)^{m-1}$ possible fluctuation-based dispersion patterns. The other calculation steps of the FDE are the same as that of DE.

C. Classifier

In related works [21], [30], [31], the authors applied several machine learning algorithms to detect whether or not the subject is MW, such as naïve Bayes (NB), linear support vector machine (SVM), RBF-kernel support vector machine (RBF-SVM), k-nearest neighbors (KNN), and random forest (RF). In this paper, we evaluate and compare the performance among the 5 most common classifiers mentioned above.

D. Evaluation Metric

The generalizability of the data is critical in scientific research. If the classifier has already seen some data from a specific subject, overfitting to the individual may happen and jeopardize the external validity of the model. To consider generalizability, cross-validation methods, including k-fold and leave-one-subject-out (k equals to the number of subjects), are required. From the experimental result from [32], the k-fold cross-validation is pessimistically biased, especially for lower values of k. However, for the higher k, k-fold cross validation suffers from high variability. In another research [33], the experimental results show that the variability of the k-fold cross validation decreases as the k increases, which is contradict to [32]. In summary, leave-one-subject-out cross-validation has lower biased than k-fold cross validation due to the larger training set and more trials to be averaged. Therefore, we apply leave-one-subject-out cross-validation to evaluate the system performance. Additionally, appropriate metrics should be used to compare the performance between the methods. In this paper, we apply three evaluation metrics, which are the weighted F1-score (F1), Cohen's Kappa coefficient (Kappa), and area under ROC curve (AUC) to compare the performance between methods.

1) Weighted F1-Score (F1): When the data number within each class is imbalanced, the classifier will tend to predict the label of the majority class, which will gain high accuracy score. Compared to accuracy, F1-score is more reliable by taking the recall and precision into consideration. The F1-score is calculated as follows:

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision}, \tag{10}$$

where

$$Recall = \frac{TP}{TP + FN},\ Precision = \frac{TP}{TP + FP},\ TP = true\ positive\ rate,\ FP = false\ positive\ rate. \tag{11}$$

Moreover, we need to consider the performance of all the classes. In this paper, we use a modified version of the F1-score, weighted F1-score:

$$F1_{weighted} = P_{MW} \times F1_{MW} + P_{non-MW} \times F1_{non-MW}, \tag{12}$$

where P_{MW} is the number of MW instances, P_{non-MW} is the number of non-MW instances, $F1_{MW}$ is the F1 score using the MW instances as the positive class, and $F1_{non-MW}$ is the F1 score using the non-MW instances as the positive class, respectively.

2) Cohen's Kappa Coefficient (Kappa, κ): Cohen's Kappa coefficient [34] stands for the agreement between two raters. It is the proportion of agreement after chance agreement is removed from consideration. The Cohen's Kappa coefficient is calculated as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, -1 \leq \kappa \leq 1, \tag{13}$$

where p_0 is the relative observed agreement among raters, and p_e is the hypothetical probability of the chance agreement. In our case, we use Kappa to measure the agreement between true labels and predicted labels. The better the detection system performs, the higher the Kappa is.

3) Area Under ROC Curve (AUC): If the classifier can output the probability of each class, then we can calculate the receiver operating characteristic (ROC) curve. In a ROC curve, the x-axis is the false positive rate, and the y-axis is the true positive rate. By calculating the area under the ROC curve (AUC), we can analyze the effectiveness of the prediction model. The chance level of AUC is 0.5, and an excellent model has an AUC close to 1.

TABLE II
HYPER-PARAMETERS OF EACH ENTROPY FEATURE

Feature Name	Hyper-parameters
MSE	sample length: 2, scale: (1,…,20)
MPE	dimension m: 3, scale: (1,…,20)
MDE	number of class: 6, dimension: 3, scale: (1,…,20)
MFDE	number of class: 6, dimension: 3, scale: (1,…,20)
Wavelet MPE	{[cA7,cD7]: {dimension: 2, scale: (1,…,20)}, [others]: {dimension: 3, scale: (1,…,20)}}
Wavelet MDE	{[cA7,cD7]: {dimension: 2, scale: (1,…,20)}, [others]: {dimension: 3, scale: (1,…,20)}}
Wavelet MFDE	{[cA7,cD7]: {dimension: 2, scale: (1,…,20)}, [others]: {dimension: 3, scale: (1,…,20)}}

TABLE III
HYPER-PARAMETERS OF EACH CLASSIFIER

	Hyper-parameters
NB	-
KNN	number of neighbors: 10, weights: 'distance', metric: 'manhattan'
Linear SVM	C: 0.01
SVM w/ rbf kernel	gamma: 1e-4, C: 5
RF	number of estimators: 700, max features: 'auto', max depth: 12

TABLE IV
PERFORMANCE EVALUATION BETWEEN EEG W/ AND W/O ICA

	EEG with ICA			EEG without ICA (Compare to w/ ICA)		
Classifier	F1	Kappa	AUC	F1	Kappa	AUC
NB	0.565	0.108	0.556	0.584 (+0.019)	0.144 (+0.036)	0.579 (+0.023)
KNN	0.563	0.100	0.569	0.602 (+0.039)	0.189 (+0.089)	0.621 (+0.052)
L-SVM	0.543	0.059	0.546	0.618 (+0.075)	0.213 (+0.154)	0.648 (+0.102)
RBF-SVM	0.581	0.135	0.602	0.654 (+0.073)	0.285 (+0.150)	0.695 (+0.093)
RF	0.582	0.148	0.617	**0.670** (+0.088)	**0.318** (+0.170)	**0.712** (+0.095)

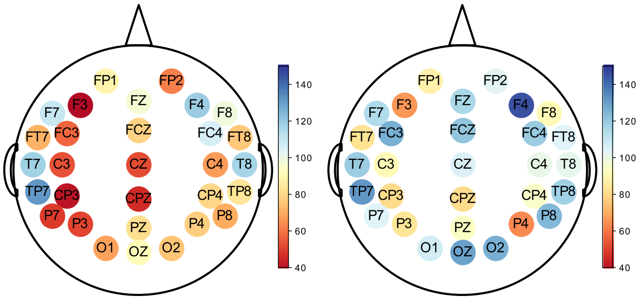

Fig. 4. Heatmaps of (a) the number of salient features among all channels on w/ICA EEG data, and (b) the number of salient features among all channels on w/o ICA EEG data. The text on each spot is the channel name.

TABLE V
SALIENT FEATURE NUMBER ON EACH CHANNEL COMPARING BETWEEN W/ AND W/O ICA EEG DATA

Ch.	w/	w/o	Ch.	w/	w/o	Ch.	w/	w/o
Fp1	**91**	90	FC4	108	**120**	CP4	82	**97**
Fp2	65	**105**	FT8	79	**106**	TP8	86	**119**
F7	111	**116**	T7	119	**120**	P7	53	**106**
F3	43	**69**	C3	57	**93**	P3	55	**86**
Fz	99	**117**	Cz	56	**107**	Pz	84	**96**
F4	120	**145**	C4	70	**103**	P4	**80**	67
F8	**101**	93	T8	**118**	104	P8	73	**125**
FT7	74	**85**	TP7	132	132	O1	71	**109**
FC3	61	**127**	CP3	45	**83**	Oz	94	**130**
FCz	81	**122**	CPz	50	**83**	O2	78	**127**

*The higher the better. Therefore, the higher number is marked in bold.

IV. EXPERIMENT RESULTS ON EEG-BASED MW DETECTION SYSTEM

To evaluate the EEG data of proposed MM-SART database, several aspects are considered. First, we analyze the overall performance of EEG data between with ICA and without ICA. Then, we evaluate the performance among channels to find out the impact of eye movement information. After that, we compare the performance in each category of features and find the most useful features for MW detection. Finally, by utilizing the feature importance metric of RF classifiers, we aim to find the most important features to detect MW.

In this experiment, we only apply 30 EEG channels without the H-EOG and the V-EOG. We first do the aforementioned preprocessing on each channel EEG, such as band-pass filter, re-reference, and labeling. After removing the subjects who have the same label in all trials, 56 subjects are included. We extract all features in Table I from the preprocessed EEG signal. As for the scale of the entropy-based features, we set it with range between 1 and 20 to get different entropy values from different scale with coarse graining as shown in Table II. We apply random search with 5-fold and 100-hyper-parameter combinations to decide the hyper-parameters of each classifier shown in Table III.

A. Performance Comparison Between EEG With and Without ICA

In this experiment, we want to observe whether eye movement information in EEG can improve the performance of the MW detection system. The EEG signal are processed with ICA (w/ ICA) and without ICA (w/o ICA). We then extract a total of 424 features per channel. Finally, the extracted features are evaluated by the aforementioned five classifiers.

From Table IV, we can observe that the best performance of both cases is acquired by applying RF classifiers as compared to other classifiers. Therefore, in the following experiment, we only focus on the performance of RF classifiers. The best F1, Kappa, and AUC of EEG w/ICA are 0.582, 0.148, and 0.617, respectively. The best F1, Kappa, and AUC of EEG w/o ICA are 0.670, 0.318, and 0.712, respectively. Therefore, EEG w/o ICA outperforms EEG w/ ICA by 0.088 for the F1, 0.170 for Kappa, and 0.095 for AUC. Therefore, eye movement information is useful when detecting MW by EEG in the MM-SART database.

Moreover, we count the number of salient features (p-value $< .05$) among all channels in both cases, as shown in Fig. 4 and Table V. While comparing w/ and w/o ICA, the number of salient features by w/o ICA data is more than that by w/ICA in most channels. Most of the features are more discriminative

TABLE VI
AUC SCORE ON EACH CHANNEL COMPARING BETWEEN W/ AND W/O ICA EEG DATA

Ch.	w/	w/o	Ch.	w/	w/o	Ch.	w/	w/o
Fp1	.541	**.602**	FC4	.518	**.590**	CP4	.523	**.576**
Fp2	.500	**.624**	FT8	.501	**.616**	TP8	.480	**.615**
F7	.524	**.597**	T7	.595	**.628**	P7	.504	**.556**
F3	.502	**.588**	C3	.533	**.596**	P3	.490	**.571**
Fz	.556	**.591**	Cz	.508	**.559**	Pz	.510	**.556**
F4	.606	**.610**	C4	.533	**.585**	P4	.546	**.564**
F8	.514	**.600**	T8	.557	**.611**	P8	.458	**.559**
FT7	.465	**.581**	TP7	.574	**.601**	O1	**.566**	.536
FC3	.505	**.557**	CP3	.479	**.557**	Oz	**.548**	.525
FCz	.519	**.587**	CPz	.537	**.559**	O2	.556	**.566**

*The higher the better. Therefore, the higher number is marked in bold.

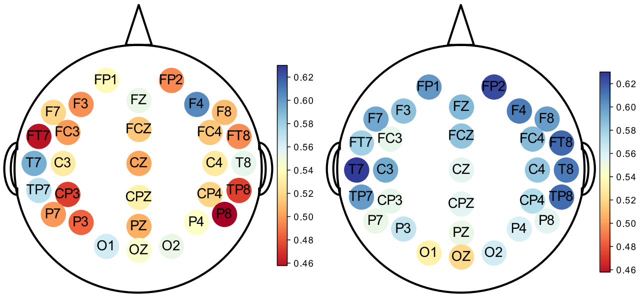

Fig. 5. Heatmaps of (a) AUC score of the RF classifier on w/ICA EEG data among all channels, and (b) AUC score of the RF classifier on w/o ICA EEG data among all channels. The text on each spot is the channel name.

because of the additional information of eye movements in EEG.

Furthermore, RF classifiers are trained for each channel to observe the AUC score of each channel. Each channel is trained and evaluated independently in both cases to observe the impact of eye movement information between channels. As shown in Fig. 5 and Table V, most of the channels have better performance on EEG w/o ICA except for channel O1 and Oz.

Also, in Fig. 5(b), we can find out that channels over the right-frontal area have better performance which has been pointed out in the previous studies [35], [36]. Moreover, the authors of [37] has pointed out that vigilance and sustained attention are impaired in subjects with right-hemisphere lesions, especially with right frontal damage [38]. Hence, from the experiment results, we verify that sustained attention is highly related to the right frontal regions (FP2), which is in line with previous clinical findings.

In addition, the channel T7, which is highly correlated to the activity of insula [39], has the second highest performance among all channels. As a core of the salience network, the insula is also in charge of the detection of emotion [40]. Since most of our self-generated thoughts involve emotional processing, it is quite straight-forward to see higher activities of insula during task-unrelated thoughts. Following the stream, the authors of [41] also proposed that the bilateral insula showed higher activity during MW, verifying the implications mentioned above. All in all, T7 is a good predictor of MW from the previous findings and further verified by the experiment result.

TABLE VII
F1/KAPPA/AUC SCORE ON SIX CATEGORIES OF FEATURES BY THE RF CLASSIFIER

F1/κ /AUC	Time	Frequency	Wavelet
Statistical	0.641/0.258/0.671	0.600/0.175/0.633	0.634/0.246/0.685
Entropy	**0.664/0.308/0.706**	0.593/0.162/0.640	0.459/0.021/0.541

TABLE VIII
F1/KAPPA/AUC ON SIX CATEGORIES OF FEATURES BY THE RF CLASSIFIER

F1/κ /AUC	Time	Frequency	Wavelet
Statistical	0.634/0.245/0.685		
Statistical + Entropy	**0.658/0.294/0.708**	0.638/0.253/0.693	0.517/0.076/0.646

In summary, we have shown that the performance of EEG w/o ICA outperforms EEG w/ICA no matter which classifier is used. Moreover, by analyzing the number of salient features and AUC of each channel, we have shown that MW is highly related to the right frontal regions and the channel T7, which is in consistent with previous clinical findings [30], [31], [33]–[37].

B. Performance Analysis on Entropy-Based Features

In this section, we analyze the effectiveness on different categories of features. Taking the aforementioned experiment results into consideration, we select RF as our classifier and apply it to EEG-without-ICA data. We train and evaluate the performance of RF classifier on each category of features, respectively. Furthermore, we concatenate the statistical features with three domains of entropy-based features and observe whether the entropy-based features can complement the statistical features and improve the performance of the RF classifier.

The experimental result of training the classifiers by each category of features respectively is shown in Table VII. The best performance among six categories of features happens when using only entropy-based features in the time domain, which has 0.664 F1-score, 0.308 Kappa, and 0.706 AUC. Moreover, the performance reached by only the entropy-based features in the time domain is close to the performance when using all features, which is 0.670 F1-score, 0.318 Kappa, and 0.712 AUC.

When combining statistical features with three domains of entropy-based features respectively, the entropy-based features in the time domain improve performance most, as shown in Table VIII. Therefore, the RF classifier can learn better with the entropy-based features in the time domain.

In summary, the extracted entropy-based features are indeed discriminative. The performance of the RF classifier can also be improved by applying entropy-based features. We have shown that the entropy-based features are suitable for EEG-based MW detection.

As a summary, we have shown that the effectiveness of overall MW detection by using EEG signals of the MM-SART database. We explore that reserving eye movement information in EEG can improve overall detection performance. Moreover, we extract entropy-based features in the time domain of EEG to complement the statistical features (i.e., mean power, power spectral density), and have shown their effectiveness in detecting MW. Finally, by analyzing the performance of different methods, we can reach 0.670 F1, 0.318 Kappa, and 0.712 AUC in the MM-SART database.

V. Channel Selection and Feature Selection on EEG-Based MW Detection System

In this section, we aim to optimize the overall system efficiency by channel selection and feature selection. The former aims to reduce the number of used EEG channels with only a slight degradation of performance. The latter aims to reduce the total number of features in the selected channels to lower the dimension of the classifier's input. Both approaches will help save the computational complexity in training the RF classifier.

A. Complexity Analysis of the RF Classifier

The RF classifier [42] is an ensemble model of decision trees. The computational complexity of building a decision tree is $O(Nkd)$, where N is the number of data, k is the number of features, and d is the number of depths of the decision tree. When building RF, two additional parameters need to be decided: the number of trees m and the number of features used to split in each node k_{sample}. Therefore, while building decision trees in RF, the computational complexity is reduced to $O(Nk_{sample}d)$. The overall computational complexity of building RF will be $O(mNk_{sample}d)$. Moreover, we have to calculate the additional computational complexity of a random selection of features at each node, which refers to $O(mkd)$. In conclusion, the final computational complexity of building the RF classifier is $O(md(Nk_{sample}+k))$. Usually, Nk_{sample} is far larger than k, so we can estimate the total computational complexity as $O(mNk_{sample}d)$.

By analyzing the computational complexity of RF, we can conclude that the training time of the RF classifier can be reduced by decreasing k_{sample}, d, and m. However, the parameters d and m are decided by the parameters searching to achieve the best performance on the data. Thus, the only method to reduce the computational complexity is to reduce the k_{sample}. In our case, k_{sample} is set to be the square root of k. In conclusion, the way to improve the training efficiency of the RF classifier is to reduce the feature dimension of the data.

B. Channel Selection and Feature Selection

As stated above, if the dimension of features can be reduced, the time complexity of training the RF classifier can be reduced, too. The number of features is $n_{channel} \times n_{feature}$, where $n_{channel}$ is the number of channels, and $n_{feature}$ is the number of features per channel. Therefore, we can first reduce the number of channels by finding the critical channels for MW detection. Then, we can perform feature selection on the critical channels to reduce the overall number of features. The following will describe the method of both reducing the number of channels and the number of features.

1) Channel Selection: To reduce the number of channels, we apply two methods to select the optimal channels, which is p-value channel selection and AUC-based channel selection. The first method is to calculate the p-value of all features of each channel. After calculating the p-value of each feature, we count the total number of salient features (p-value $< .05$). We then sort the channel with the number of salient features and select the top-k channels. This method has the advantage of fast calculation without training extra classifiers. However, the salient features are not equal to the significant features, which are decided by the feature importance of the RF classifier.

In contrast, we apply the other method to select the optimal channels. We first train the RF classifiers per channel. We evaluate the performance of each trained classifier and sort the channel by its AUC. Finally, we can select top-k channels as the candidates to train the final classifier.

Although this method requires training additional classifiers, it has the advantage of having better performance. If the specific channel itself can perform well by only its own features, we can infer that by selecting the optimal number of such channels, we can find the best trade-off between performance and computational complexity.

2) Feature Selection: After reducing the number of channels, we further lower the number of features in each channel to reduce the time complexity during training. In [43], the feature selection method has been shown to improve learning performance, increase computational efficiency, decrease memory storage, and build better generalization models. Therefore, we propose a specific feature selection method for RF classifiers: correlation importance feature elimination (CIFE). The CIFE contains two steps: unsupervised correlation clustering and supervised importance rejection.

First, we calculate the correlation between pairs of features with the following equations,

$$\rho\ (\mathrm{X},\mathrm{Y}) = \frac{\sum_i (x_i-\bar{x}_i)(y_i-\bar{y}_i)}{\sqrt{\sum_i (x_i-\bar{x}_i)^2 \sum_j (y_j-\bar{y}_j)^2}}\ , \tag{14}$$

where X,Y are two different features, and i and j are the ith and jth data respectively. After that, features with correlations higher than ρ_{thres} are clustered. Among each cluster, only one feature will be selected as the representative. Therefore, after correlation clustering, the number of remaining features will equal to the number of clusters. This process can eliminate features that are too similar to highly reduce the number of features.

The second step is to perform supervised importance rejection. Different from unsupervised correlation clustering, supervised importance rejection requires labels to pretrain a RF classifier. After training the RF classifier, top-k remaining features are selected according to the feature importance of the RF classifier. Therefore, the final number of features is k. While comparing with recursive feature elimination (RFE) [44], CIFE is a one-path method which is not necessary to recursively train the classifier. Therefore, the training efficiency of CIFE will be higher than RFE during the selection process. Moreover, while comparing with importance feature elimination (IFE) [45] using the feature importance score of the RF, CIFE eliminates the similar features in the first step which can prevent selecting redundant features.

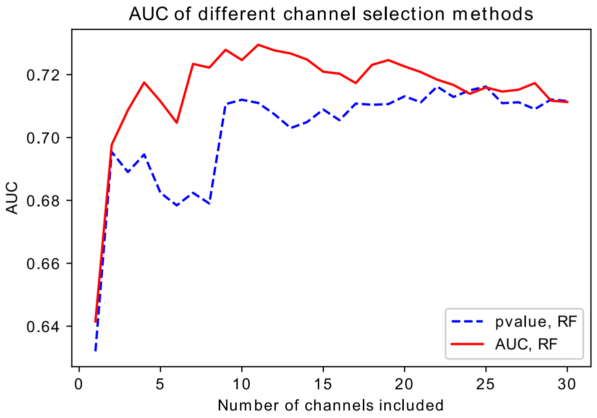

Fig. 6. AUC of the *p*-value channel selection and the AUC-based channel selection.

Therefore, CIFE can reach better AUC with fewer features than IFE.

C. Experimental Results

To show the improvement of the efficiency in our proposed method, we verify the improvements from the channel selection and feature selection, respectively. First, we evaluate the improvement of channel selection and choose an optimal number of channels. Second, we evaluate the improvement of our proposed CIFE. We implement our experiment on the previously selected channels. Also, we compare different methods by AUC and training time.

The processing of the EEG signals is the same as the previous section. However, considering the issue of efficiency, we eliminate MSE features due to their high time complexity [19].

1) Experiments on Channel Selection: In order to lower both training and inference complexity of the overall MW detection system, channel selection is necessary. In this experiment, we want to analyze the efficiency of the two channel selection methods, *p*-value channel selection, and AUC-based channel selection. We first sort the thirty channels according to the number of salient features and AUC, respectively. We then evaluate the performance of each method by adding one channel at a time. Finally, the AUC is used to evaluate the performance.

As shown in Fig. 6, we can see that with only one channel, the performance can only reach 0.61$\sim$0.65 AUC. By adding one more channel, the performance of each method reaches close to 0.7 AUC. When comparing two methods, we can see that the performance of AUC-based selection is slightly better than that of *p*-value selection. As a result, we select two channels to get the best trade-off between computational complexity and performance. The detailed comparison is shown in Table IX. In conclusion, by selecting two critical channels, we lose only 0.016 in AUC but decrease 44.16% of training time in RF classifiers compared to the original 30 channels.

2) Experiments on Feature Selection: In this section, we aim to further improve the system efficiency with feature selection methods. Therefore, we compare three methods: correlation importance feature elimination (CIFE), importance feature elimination (IFE), and recursive feature elimination (RFE). In the following experiment, we use two critical channels mentioned in the previous section (V.C.1) (i.e., T7 and Fp2). Therefore, the original dimension of the features is 808 (404 features/channel × 2 channels). Moreover, the ρ_{thres} of CIFE is set to 0.9.

From Fig. 7 and Table X, when the selected number of features is small, the best feature selection method is RFE, and the worst method is IFE. Moreover, the AUC score of CIFE with 14 features is comparable to the AUC score without feature selection, and the AUC score of RFE with 11 features is comparable to the AUC score without feature selection. However, when the number of features increases, RFE and CIFE converge to almost the same AUC score. In contrast, CIFE performs closely to RFE, but the computational time of CIFE is only 31% of RFE. Therefore, when considering performance and training efficiency, CIFE is the best choice among the three methods.

TABLE IX
PERFORMANCE (AUC) AND TIME COMPARISON AMONG BOTH CHANNEL SELECTION METHODS

	Performance (AUC)	Time (sec)
Original (30 channels)	0.712	236.23
P-value channel selection (2 channels, F4 and TP7)	0.687 (-0.025)	132.93 (-43.72%)
AUC-based channel selection (2 channels, T7 and FP2)	0.696 (-0.016)	131.90 (-44.16%)

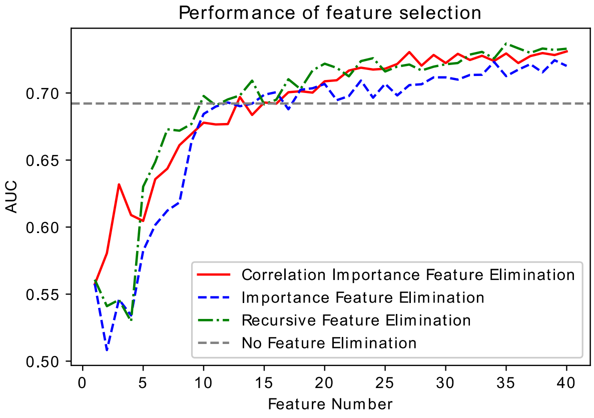

Fig. 7. Feature selection performance evaluation. The red line represents no feature elimination, that is to say, we train the RF classifier with 808 features.

TABLE X
PERFORMANCE AND TIME COMPARISON AMONG FEATURE SELECTION

Feature number		AUC (%)	Overall Time (sec)	Selection Time (sec)
808		69.4	177.02	
40	RFE	73.2 (+3.8)	424.76	308.44
	IFE	71.1 (+1.7)	157.97	41.15
	CIFE	**72.5 (+3.1)**	**164.58 (-61.3%*)**	**45.65 (-85.2%*)**
11	RFE	69.8	441.72	321.81
15	IFE	69.5	161.46	41.79
14	CIFE	69.6	172.18	47.14

*Comparing to RFE with 40 features.

TABLE XI
SELECTED FEATURES AFTER DIFFERENT FEATURE SELECTION METHODS

Method (Number)	Feature Name
RFE (11)	T7_PSD_beta, T7_MDE-20, FP2_Mean, FP2_PSD_beta, FP2_PSD_gamma, FP2_cD4-WL-MeanPower, FP2_cD4-WL-RAM, FP2_MFDE-17, FP2_SpecEnt_beta, FP2_SpecEnt_gamma, FP2_cD7-WL-Ent
IFE (15)	T7_PSD_beta, T7_MFDE-19, T7_MFDE-20, T7_MDE-18, T7_MDE-20, T7_WL-MFDE-cD7-14, FP2_PSD_beta, FP2_PSD_gamma, FP2_cD4-WL-MeanPower, FP2_cD4-WL-STD, FP2_cD4-WL-RAM, FP2_SpecEnt_beta, FP2_SpecEnt_gamma, FP2_cD7-WL-Ent, FP2_cD7-WL-SpecEnt
CIFE (14)	T7_FirstDiff, T7_HjComp, T7_PSD_beta, T7_PSD_gamma, T7_MFDE-1, T7_WL-MFDE-cD7-14, FP2_Mean, FP2_FirstDiff, FP2_PSD_theta, FP2_PSD_beta, FP2_PSD_gamma, FP2_MPE-1, FP2_cD7-WL-Ent, FP2_cD7-WL-SpecEnt

As shown in Table XI, we list the optimal features selected by these three methods. T7_PSD_beta, FP2_PSD_beta, FP2_PSD_gamma, and FP2_cD7-WL-Ent are chosen by all three methods. Moreover, when observing the category of the selected features, we discover that the selected features belong to different categories. Therefore, we conclude that the complementarity between each category of features can help improve the overall performance of the MW detection system based on EEG.

VI. CONCLUSION

In this paper, we present a multi-modality sustained attention to response task (MM-SART) database. We also propose a framework to detect MW based on the EEG signals collected in the MM-SART database. In our framework, entropy-based features can complement traditional EEG features and therefore improve the performance of the overall system. Moreover, by selecting the best two critical channels, T7 and Fp2, and applying correlation importance feature elimination framework for RF classifiers, we can improve the performance and computational efficiency of the MW detection system based on EEG. The final AUC score of our framework is 0.725 with two channels and forty features in total. To this end, detecting MW is critical in our daily lives, as MW can lead to negative effects on emotions, and also influence our learning efficiency and driving safety. We hence propose a framework that can detect MW and hopefully this can be utilized in the educational scenarios.

APPENDIX

SUPPLEMENT OF THE MM-SART DATABASE

A. Behavioral Performance

We briefly summarize the behavioral performance here. In the current study, successful stop rate (withhold response while seeing target letter “C”) is 73.18%. In the reaction time (RT) and reaction time coefficient of variation (RTCV) analysis, we compare the performance of the 5 trials preceding the target letter “C” and the 5 trials preceding the probe across participants. Precisely, we extract a 10-second time window, which is the same as the classification analysis, to compare the RT and RTCV in objective/subjective focused and wandering states, as shown in Table XII. Significant slower RTs are found in successful stop conditions compared to fail-to-stop conditions ($p < .001$). Additionally, a trend for smaller RTCVs in the successful stop condition compared to those of the fail-to-stop condition is found ($p = .065$). On the contrary, there is no significant difference in RT when comparing subjectively rating focused compared to rating wandering ($p = .199$). However, a significant larger RTCV is found in the rating wandering condition compared to that of the rating focused condition ($p < 0.001$).

TABLE XII
THE BEHAVIORAL PERFORMANCE IN THE TASK

Mean (STD)	successful stop	fail-to-stop	rating focused	rating wandering
RT (ms)	412.16 (63.37)	375.8 (58.46)	386.07 (78.79)	395.67 (90.77)
RTCV	0.19 (0.09)	0.21 (0.1)	0.21 (0.09)	0.35 (0.2)

ACKNOWLEDGMENT

The authors wish to thank the participants that engaged in the data collection of MM-SART

REFERENCES

[1] J. Smallwood et al., “Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance,” *Psychon. Bull. Rev.*, vol. 14, pp. 230–236, Apr. 2007.

[2] R. Bixler and S. D’Mello, “Toward fully automated person-independent detection of mind wandering,” in *User Modeling, Adaptation, and Personalization*, Cham, Switzerland: Springer, 2014.

[3] K. K. Szpunar et al., “Mind wandering and education: From the classroom to online learning,” *Frontiers Psychol.*, vol. 4, pp. 1–7, 2013.

[4] G. Pepin et al., “Impact of mind-wandering on visual information processing while driving: An electrophysiological study,” *Appl. Cogn. Psychol.*, vol. 35, pp. 508–516, 2021.

[5] M. A. Killingsworth et al., “A wandering mind is an unhappy mind,” *Science*, vol. 330, p. 932, 2010.

[6] M. Chen et al., “Negative information measurement at AI edge: A new perspective for mental health monitoring,” *ACM Trans. Internet Technol.*, vol. 22, no. 3, pp. 1–6, Jan. 2022.

[7] T. Manly et al., “The absent mind: Further investigations of sustained attention to response,” *Neuropsychologia*, vol. 37, pp. 661–670, 1999.

[8] R. Bixler et al., “Automatic detection of mind wandering during reading using gaze and physiology,” in *Proc. ACM Int. Conf. Multimodal Interaction*, New York, NY, USA, 2015, pp. 299–306.

[9] C. L. Baldwin et al., “Detecting and quantifying mind wandering during simulated driving,” *Frontiers Human Neurosci.*, vol. 11, 2017, Art. no. 406.

[10] E. Barron et al., “Absorbed in thought: The effect of mind wandering on the processing of relevant and irrelevant events,” *Psychol. Sci.*, vol. 22, pp. 596–601, 2011.

[11] C. Y. Jin et al., “Predicting task-general mind-wandering with EEG,” *Cogn., Affect., Behav. Neurosci.*, vol. 19, pp. 1059–1073, 2019.

[12] G. Gupta, S. Pequito, and P. Bogdan, “Re-thinking EEG-based non-invasive brain interfaces: Modeling and analysis,” in *Proc. IEEE/ACM 9th Int. Conf. Cyber-Phys. Syst.*, 2018, pp. 275–286.

[13] Y. Xue, S. Rodriguez, and P. Bogdan, “A spatio-temporal fractal model for a CPS approach to brain-machine-body interfaces,” in *Proc. IEEE DATE*, 2016, pp. 642–647.

[14] G. Gupta et al., “Dealing with unknown unknowns: Identification and selection of minimal sensing for fractional dynamics with unknown inputs,” in *Proc. Annu. Amer. Control Conf.*, 2018, pp. 2814–2820.

[15] F. C. Morabito et al., “Multivariate multi-scale permutation entropy for complexity analysis of Alzheimer’s disease EEG,” *Entropy*, vol. 14, pp. 1186–1202, 2012.

[16] F. Ghassemi et al., “Using non-linear features of EEG for ADHD/normal participants’ classification,” *Procedia-Social Behav. Sci.*, vol. 32, pp. 148–152, 2012.

[17] C. Goh et al., “Comparison of fractal dimension algorithms for the computation of EEG biomarkers for dementia,” in *Proc. 2nd Int. Conf. Comput. Intell. Med. Healthcare*, Lisbon, Portugal, 2005.

[18] K. Tung et al., “Entropy-assisted multi-modal emotion recognition framework based on physiological signals,” in *Proc. IEEE-EMBS Conf. Biomed. Eng. Sci.*, 2018, pp. 22–26.

[19] W. Aziz and M. Arif, “Multiscale permutation entropy of physiological time series,” in *Proc. Pakistan Sect. Multitopic Conf.*, 2005, pp. 1–6.

[20] M. Costa et al., “Multiscale entropy analysis of complex physiologic time series,” *Phys. Rev. Lett.*, vol. 89, 2002, Art. no. 068102.

[21] R. Bixler and S. D’Mello, “Automatic gaze-based user-independent detection of mind wandering during computerized reading,” *User Model. User-Adapted Interaction*, vol. 26, no. 1, pp. 33–68, Mar. 2016.

[22] G. Manis, “Fast computation of approximate entropy,” *Comput. Methods Programs Biomed.*, vol. 91, pp. 48–54, 2008.

[23] “MMSART database.” [Online]. Available: http://mmsart.ee.ntu.edu.tw

[24] M. Mittner et al., “When the brain takes a break: A model-based analysis of mind wandering,” *J. Neurosci.*, vol. 34, pp. 16286–16295, 2014.

[25] N. Hu et al., “Different efficiencies of attentional orienting in different wandering minds,” *Consciousness Cogn.*, vol. 21, pp. 139–148, 2012.

[26] C. Kalina et al., “Experience sampling during fMRI reveals default network and executive system contributions to mind wandering,” *Proc. Nat. Acad. Sci.*, vol. 106, pp. 8719–8724, May 2009.

[27] A. Hyvärinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Netw.*, vol. 13, pp. 411–430, 2000.

[28] H. Azami, M. Rostaghi, D. Abásolo, and J. Escudero, “Refined composite multiscale dispersion entropy and its application to biomedical signals,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2872–2879, Dec. 2017.

[29] H. Azami et al., “Multiscale fluctuation-based dispersion entropy and its applications to neurological diseases,” 2019, *arXiv:1902.10825*.

[30] S. Hutt et al., “The eyes have it: Gaze-based detection of mind wandering during learning with an intelligent tutoring system,” *Int. Educ. Data Mining Soc.*, 2016.

[31] M. Faber et al., “An automated behavioral measure of mind wandering during computerized reading,” *Behav. Res. Methods*, vol. 50, pp. 134–150, Feb. 2018.

[32] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, pp. 338–345.

[33] Y. Bengio and Y. Grandvalet, “No unbiased estimator of the variance of K-fold cross-validation,” *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, 2004.

[34] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, pp. 37–46, 1960.

[35] I. McGilchrist, “Reciprocal organization of the cerebral hemispheres,” *Dialogues Clin. Neurosci.*, vol. 12, pp. 503–515, 2010.

[36] M. Sarter et al., “The cognitive neuroscience of sustained attention: Where top-down meets bottom-up,” *Brain Res. Rev.*, vol. 35, pp. 146–160, Apr. 2001.

[37] L. Rueckert et al., “Sustained attention deficits in patients with right frontal lesions,” *Neuropsychologia*, vol. 34, pp. 953–963, Oct. 1996.

[38] A. J. Wilkins et al., “Frontal lesions and sustained attention,” *Neuropsychologia*, vol. 25, pp. 359–365, Jan. 1987.

[39] A. R. Hidalgo-Muñoz et al., “Spectral turbulence measuring as feature extraction method from EEG on affective computing,” *Biomed. Signal Process. Control*, vol. 8, pp. 945–950, Nov. 2013.

[40] J. Markovic et al., “Tuning to the significant: Neural and genetic processes underlying affective enhancement of visual perception and memory,” *Behav. Brain Res.*, vol. 259, pp. 229–241, Feb. 2014.

[41] M. F. Mason et al., “Wandering minds: The default network and stimulus-independent thought,” *Science*, vol. 315, pp. 393–395, Jan. 2007.

[42] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[43] J. Li et al., “Feature selection: A data perspective,” *ACM Comput. Surveys*, vol. 50, pp. 1–45, 2018.

[44] P. M. Granitto et al., “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemometrics Intell. Lab. Syst.*, vol. 83, pp. 83–90, 2006.

[45] M. Saarela and S. Jauhiainen, “Comparison of feature importance measures as explanations for classification models,” *SN Appl. Sci.*, vol. 3, 2021, Art. no. 272.