Computers and Electrical Engineering 122 (2025) 110005

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

MSSTNet: A multi-stream time-distributed spatio-temporal deep learning model to detect mind wandering from electroencephalogram signals

Subrata Pain [a],*, Subhrasankar Chatterjee [b], Monalisa Sarma [c], Debasis Samanta [b]

[a] *Advanced Technology Development Center, Indian Institute of Technology Kharagpur, India*
[b] *Department of Computer Science & Engineering, Indian Institute of Technology Kharagpur, India*
[c] *Subir Chowdhury School of Quality and Reliability, Indian Institute of Technology Kharagpur, India*

ARTICLE INFO

Keywords:
Mind wandering
Electroencephalogram signal processing
Spatio-temporal EEG
Deep convolutional-recurrent neural networks
Multi-stream models

ABSTRACT

The automated detection of mind-wandering (MW) and associated attention lapses through Electroencephalogram (EEG) signals holds significant potential for practical applications. Traditional handcrafted features have been proven inadequate in capturing the spatially and temporally distributed patterns of MW-EEG signals, limiting the effectiveness of Machine Learning-based detection models. This study proposes a multi-stream spatio-temporal Deep Learning network (MSSTNet) for automated detection of MW episodes. In this approach, EEG data from individual frequency bands, extracted directly from raw EEG signals, are fed into distinct feature extraction blocks embedded within each stream of the MSSTNet architecture. These feature extraction blocks comprise time-distributed Convolutional Neural Network—Long Short-Term Memory (CNN-LSTM) modules, enabling the model to capture fine-grained spatio-temporal features across different frequency bands. The CNN layers extract spatial and short-term temporal features, and the LSTM units capture the long-term temporal evolution of these features, which is critical for recognizing prolonged MW episodes. The features extracted from each frequency band are subsequently sent to fully connected layers for MW detection. The MSSTNet model is validated on a publicly available MW and focus dataset, achieving a mixed-subject classification accuracy of 95.07%, substantially outperforming baseline models. Furthermore, the intra-subject and cross-subject classification accuracies of 94.48% and 83.13%, respectively, demonstrate the robustness and generalizability of the proposed model. The band-wise analysis reveals that the Beta band exhibits the most pronounced alterations due to MW onset. MSSTNet's capacity to capture subtle spatio-temporal patterns across frequency bands underscores its efficacy as an MW detection framework, with promising scope for broader EEG-based applications.

1. Introduction

The human brain frequently shifts focus away from ongoing tasks towards self-centered thoughts, a phenomenon known as mind wandering (MW). MW is characterized by a drift from task-related attention to non-task-related, personality-inward thoughts. MW occurs more frequently during low cognitive demand tasks, when the human mind and body are relatively relaxed. The prevalence of such stimulus-independent thought episodes is significant, accounting for approximately 46.9% of daily activities [1].

* Corresponding author.
E-mail addresses: subratabankata@gmail.com (S. Pain), sc1933@cse.jgec.ac.in (S. Chatterjee), monalisa@iitkgp.ac.in (M. Sarma), dsamanta@iitkgp.ac.in (D. Samanta).

https://doi.org/10.1016/j.compeleceng.2024.110005
Received 23 May 2024; Received in revised form 28 November 2024; Accepted 8 December 2024
Available online 17 December 2024

This redirection of focus has a detrimental impact on performance across various tasks, including real-time driving, listening to live lectures, and working memory tasks [2]. For efficient monitoring of human performance and avoiding potentially hazardous situations like accidents, an automated, real-time, and efficient framework for detecting MW-induced attention loss is essential.

Electroencephalogram (EEG)-based brain signal processing has recently emerged as a powerful tool for analyzing the neurophysiological signatures of MW episodes. Early research in this area has investigated MW-induced attenuation of multiple Event-Related Potential (ERP) components within MW-EEG signals, observing notable suppression in the N1, P1, and P300 components [3]. Multiple studies have also explored the effects of MW onset on the oscillatory activity of various EEG frequency bands [2–6]. In [4], the MW onset during a breath counting task has been investigated, and a prominent increase of the Theta power is observed in all electrode positions. In [5], reduced Alpha-activity is identified as a biomarker of MW while performing a Stroop task. In contrast, an increase in Alpha power during MW periods is reported in simulated driving [3] and a switching task [6]. These findings collectively indicate the influence of MW on EEG frequency bands, although the results remain inconsistent across different tasks.

The lack of generalized MW biomarkers continues to hinder the performance of Machine Learning (ML)-based MW detection models that utilize these ERP patterns and band powers as features. In [7], a set of temporal features, including band powers and ERP components, are extracted, and classification is performed using Support Vector Machine (SVM). Similarly, in [8] Independent Component Analysis (ICA) is performed on Alpha band EEG data, and the power of the resulting ICs is utilized as features for a non-linear SVM classifier. In another study [9], Beta and Gamma band powers are utilized for classification using an RBF-SVM (Radial Basis Function Kernel SVM). However, the accuracy of these models has remained relatively low and varied significantly depending on participants. To determine the most affected frequency band during MW onset, Common Spatial Patterns (CSPs) are extracted from individual frequency bands, and ML-based classification is applied in [2]. However, significant subject-wise variability is observed, and the overall effect of MW on EEG frequency bands remains inconclusive.

In contrast to the handcrafted feature-based ML models, Deep Learning (DL) techniques offer automated feature extraction and classification for various EEG tasks, including MW detection. Convolutional filters have been proven effective in learning spatial and local temporal features [10,11] from EEG signals. On the other hand, recurrent models are adept at extracting long-term contextual patterns from temporal EEG [12,13]. In line with these techniques, in [14], a CNN-based MW detection model utilizing raw EEG data ($0.1-40$ Hz) is proposed. In this study [14], the temporal and spatial features are extracted using 1-D convolution over both dimensions. Again, in [15], a CNN-based model is proposed to detect MW across various tasks. Instead of using the raw EEG data, spectrograms of the EEG signals are extracted, and different CNN models are trained individually for each feature. The individual models' outcomes are combined to get the final classification. However, the detection performance of MW across various tasks [15] is severely inadequate, highlighting the challenges of obtaining task-independent MW features. Moreover, a Recurrent Neural Network (RNN)-based model is proposed in [16] to extract the evolving temporal dynamics in prolonged MW episodes. Here, the spatial features extracted by pre-trained convolutional filters are sent to the RNN model for contextual feature extraction. However, no time-sequencing of the input EEG epochs is performed while extracting the convolutional features. So, the purpose of obtaining the temporal development of MW features remains unaccomplished when these non-sequential features are utilized as input to the RNN model.

Despite the superior performance of DL-based feature extractors compared to handcrafted techniques, a thorough analysis reveals the following limitations:

1. As previously noted, distinct EEG frequency bands play a crucial role in delineating the effects of MW onset [2,4,5]. However, most DL models train convolutional filters on the full spectrum (0.1–40 Hz) of MW-EEG signals. Consequently, the trained filters are inadequate to capture the precise and fine-grained MW properties existing in each frequency band.
2. During prolonged MW episodes, EEG signals exhibit a temporal evolution of neural activity [17,18]. While some studies apply 1-D convolution over time to capture short-term temporal features [14,15], the long-term progression of MW-EEG features remains insufficiently characterized due to the restricted receptive field of convolutional filters.

To alleviate the aforementioned issues, this study proposes a multi-stream spatio-temporal deep learning model (MSSTNet) for efficient MW detection. Each stream of the proposed model consists of a feature extraction block that receives the signals from the distinct frequency bands and performs spatio-temporal feature extraction. Consequently, the learned filters in each stream mine subtle and fine-grained information from the individual bands. The proposed feature extraction block consists of multiple convolution layers, followed by a long short-term memory (LSTM) network. In contrast to other models that extract convolutional features from the entire EEG epoch without preserving the time sequence [19,20], the proposed model adopts a different approach. Here, the extracted EEG epochs are fragmented into multiple smaller intervals and fed to the convolutional blocks in a time-distributed manner. The spatial and short-term temporal patterns are extracted from the epoch fragments by 1-D spatial and temporal filters. Extracted features are then sequentially fed to the LSTM network that precisely captures the long-term evolution of MW-EEG properties. Finally, the individual band features obtained from the streams are concatenated to form a single feature vector, and the classification between MW and Focus states is performed through a series of fully connected layers.

The training of separate spatio-temporal feature extractors for individual EEG frequency bands rather than using a single extractor for the entire spectrum has some potential advantages. Each frequency band, especially Theta, Alpha, Beta, and Gamma, contains unique features that are highly relevant to MW detection. When a single extractor is trained on the full spectrum (0.1–40 Hz), the convolutional filters are forced to capture a broad set of features, potentially diluting or ignoring the specific, fine-grained information embedded within each band. This limitation can reduce the model's ability to detect the nuanced, band-specific patterns crucial for identifying MW onset. In contrast, training separate feature extractors for each frequency band allows for targeted filters optimized to capture the distinct spatio-temporal characteristics of each band. This focused learning enhances the model's ability to detect subtle, discriminatory features. By subsequently combining the extracted features from all bands, the model gains a more comprehensive representation of the EEG data, leading to improved generalization and potentially higher classification performance. The efficacy of this proposed approach is quantitatively validated through extensive experiments.

The main contributions of this research include the following:

- A multi-stream DL model that is capable of extracting adequate discriminatory features from individual frequency bands of MW-EEG data.
- An efficient spatio-temporal feature extractor that precisely captures the spatial-, short-, and long-term evolving temporal features from each frequency band's signals.
- Identification of the EEG frequency band that is most affected by MW-onset by performing individual band-wise performance analysis.

2. Related work

This section includes a detailed discussion of the ongoing research on identifying discriminatory EEG biomarkers of MW. Further, the techniques for EEG signal-based automated detection of MW are discussed.

2.1. Identification of MW using biomarkers from EEG frequency bands

The alteration of EEG frequency band oscillations induced by MW has been initially investigated in [4], using resting state EEG during a breath-counting task. Through the comparison of absolute log-spectral power across all frequency bands, it is seen that Theta band power exhibits the most significant alteration, namely an increase in Theta power across all electrodes during MW. Changes in other bands are less pronounced and spatially localized. In contrast, in another study [5], a reduced Alpha and Theta activity during MW is reported during a Stroop task. Further, an enhanced Alpha power is reported in MW episodes during a simulated driving task [3] and a font-color identification task [21]. The most recent analysis [6] also finds increased Alpha power for MW during a number detection task. While some studies identify enhanced Alpha power as a biomarker of MW, results exhibit inconsistency across investigations. Additionally, the effects of MW on frequency bands other than Alpha are not well-established. This study aims to explore the impact of MW on all frequency bands through a series of experiments.

2.2. Automated detection of MW and focus using ML and DL models

Researchers have traditionally used either handcrafted feature extraction, followed by ML models, or DL models with raw or handcrafted EEG data to classify EEG signals. ML-based MW detection algorithms are reviewed first. We also briefly describe generic DL-based EEG classification methods, as our work focuses on DL-based detection. Finally, we discuss DL-based MW detection approaches, model designs, benefits, and drawbacks.

2.2.1. ML-based MW detection using handcrafted features

In [7], MW-related attention loss during sustained attention and visual search tasks has been investigated. Known neural correlates, such as Theta and Alpha band power, coherence, and P1, N1, and P3 responses, have been considered as features, and an ML-based classification has been performed using these features. Depending on the participants, the SVM classifier used in this study predicts MW with 50% to 85% accuracy. To detect MW during live lectures, a CSP-based data-driven approach is followed in [2]. Classification accuracy ranging from 67% to 100% is obtained for different subjects. Notably, in this study, the frequency band most affected by MW events remains unidentified [2], as features providing the highest performance for different subjects vary.

2.2.2. Deep learning-based methods for EEG classification

Recently, DL models have been widely employed to mine complicated EEG patterns. CNNs can extract and classify localized and stationary spatial features from EEG data [22–25]. There are two ways to use CNN for EEG categorization. In the first technique, the raw data is transformed into a new representational feature space using different signal processing techniques, and the extracted features are fed to CNN for mining more complex patterns. In the second technique, the CNN model extracts features directly from the raw EEG data. Both methods are briefly discussed next.

In [25], the raw EEG data is first transformed into spectrogram images using Short Time Fourier Transform and Mel scale spectrogram methods, and the images are then classified using CNN-based models. Here, two pre-trained image classification models, Google Network (GoogLeNet) and Residual Network-34 (ResNet-34) are used for emotion detection from EEG data. The best performance is achieved with a fine-tuned GoogLeNet model using STFT spectrogram images as input. Similarly, in [26], EEG data is first decomposed into components using discrete wavelet transform, empirical mode decomposition, and empirical wavelet transformation. Each component is then fed to CNNs with identical architectures, and the final classification output is obtained by using ensemble methods. In another study [27], Epileptic seizure EEG signals are decomposed into wavelet coefficients using multi-scale wavelet transform, and each coefficient is processed by attention-based CNN blocks for feature extraction. When validated on two EEG datasets, the framework proves to be efficient for multi-class Epilepsy detection.

More recently, a few advanced DL models have been proposed for EEG classification. In this regard, EEGNet [10] and HTNet [11] models have become popular as generalized EEG decoders. The EEGNet model consists of 3 convolutions and 1 dense layer, along with dropout and average pooling layers after each convolution layer. In EEGNet, depthwise and separable convolution layers are introduced to extract patterns from EEG sequences, yielding robust performance across different EEG paradigms. Further, HTNet extends EEGNet with a Hilbert transform and electrode-level projection operations for generalized performance. In addition, the authors of [28] have employed a compact and interpretable CNN model (ICNN) that uses separable convolution for efficient driver fatigue detection. To extract the spatio-temporal characteristics, the separable convolution architecture employs two convolution operations. The depthwise convolution is applied to each individual electrode's signal independently, followed by a pointwise convolution. Further, a novel interpretation method is proposed to investigate the neurological significance of ICNN extracted features. In [29], a 1-D CNN and a 1-D capsule network are proposed for the classification of preictal and interictal EEG signals. In addition to convolutional layers, this approach includes specialized layers, such as the primary capsule layer and digitcaps layer, to achieve improved feature representation by capturing the temporal hierarchy within the EEG time sequence. Although the model is highly efficient in detecting epileptic seizures, one drawback of this strategy is its limited ability to process data from only one EEG channel. Therefore, this model is unsuitable for extracting features from multiple EEG channels. None of these models utilize recurrent layers. Due to the limited kernel size in these models, the convolutional filters can only learn features with local receptive fields, potentially missing the long-term evolution of the EEG features.

To capture long-term evolutionary patterns in EEG sequences, RNN models, particularly GRUs [30] and LSTMs [31], are employed for EEG decoding. In [30], a three-layer GRU with 100 units, followed by dense layers, is proposed for sleep staging. While GRUs are computationally efficient compared to LSTMs, they may struggle with long-term dependencies due to the simpler gating mechanism and absence of the memory cell. LSTMs are better at handling these dependencies. In [31], dimensionality-reduced EEG streams are processed by LSTMs and dense layers for Schizophrenia detection. However, due to their sequential processing, LSTM-based models may also experience vanishing gradient issues during training for lengthy time-series. To address this, the Transformer models, equipped with self-attention and parallel processing, are also being used for EEG [32]. In [32], an improved time-frequency transformer (TFormer) is proposed to capture global time-frequency patterns. The TFormer model comprises three components: convolutional stems for input embedding, time-frequency multi-head cross-attention, and self-attention to extract time-frequency features. Using both time and frequency representation of EEG, the model exhibits superior performance in driver's fatigue detection. Because of their success, it is worthy to investigate the model's performance in other EEG applications.

Apart from the aforementioned backpropagation-based DL models, most recently, randomized neural networks, such as Random Vector Functional Link (RVFL) networks, have also emerged for EEG decoding [33,34]. In randomized networks, the weights of the hidden nodes are randomly initialized and kept fixed during training, while the weights of the output nodes are updated through training, not via backpropagation but by using closed form solutions. When applied to small volumes of EEG datasets, this approach has some potential benefits over traditional DL models, such as less overfitting, a lower number of parameters, and faster training. In [33,34], two advanced RVFL models, namely spectral-ensemble deep RVFL (SedRVFL) and random forest-based feature selection, global output, weighting, and entropy-based dynamic ensemble RVFL (FGloWD-edRVFL), have been proposed. The SedRVFL model employs a feature-refining block, dynamic direct links from input to output, and ensemble methods to improve the low feature-learning capability of randomized networks. In contrast, the FGloWD-edRVFL model uses a global output layer, a specialized entropy-based ensemble method, and a random forest-based feature selection to discard inferior features from the randomized feature space. These pioneering models exhibit superior performance in driver's fatigue detection, and there is a need to explore these models' generalization in different EEG paradigms.

2.2.3. MW detection using deep learning

Following the aforementioned techniques, in [14], a CNN-based model consisting of four convolution and two dense layers is proposed for MW and focus detection. In [16], 2-D topographic maps are created from raw EEG data, and the spatial features are extracted using a pre-trained residual network-50 (ResNet-50) model by exploiting transfer learning. Extracted spatial features are subsequently fed to the RNN model. The major drawback of this approach is that the spatial features do not have any temporal order, so the RNN layers are unable to extract the time-evolving nature of MW-EEG features. Moreover, the ResNet-50 model, pre-trained on a set of natural images and is inefficient for EEG data since the natural images are spatially continuous, whereas the EEG activations are spatially discrete. Again, in [15] multiple manually extracted features are fed to multiple CNN models for further pattern mining. The output of the different models is then combined using a meta-learner to produce the final classification. Noteworthy, in [15], cross-task MW detection was performed, i.e., training a MW detection model for a particular task and testing the model by detecting MW events for a different task. However, the obtained detection performance has been unsatisfactory. In all of the aforementioned CNN-based models, the learned filters attempted to extract features from EEG data containing the entire frequency range. So, these models are inefficient at capturing subtle discriminatory features from individual frequency bands. In this study, by employing separate feature extractors for individual bands, specialized filters, trained to extract fine-grained features from each frequency band, are learned.

3. Methods

Fig. 1 depicts a conceptual overview of the proposed framework. Firstly, the MW and focus epochs are extracted from the continuous EEG recordings. Then band-pass filters are applied to the extracted epochs to obtain the individual frequency bands' data. After that, individual frequency bands' data is sent to the streams of the MSSTNet model for spatio-temporal feature extraction. The feature extractor block consists of time-distributed CNN-LSTM layers that extract fine-grained features from each frequency band. The extracted features are concatenated to form a single feature vector, and lastly, the classification is performed by a series of dense layers.

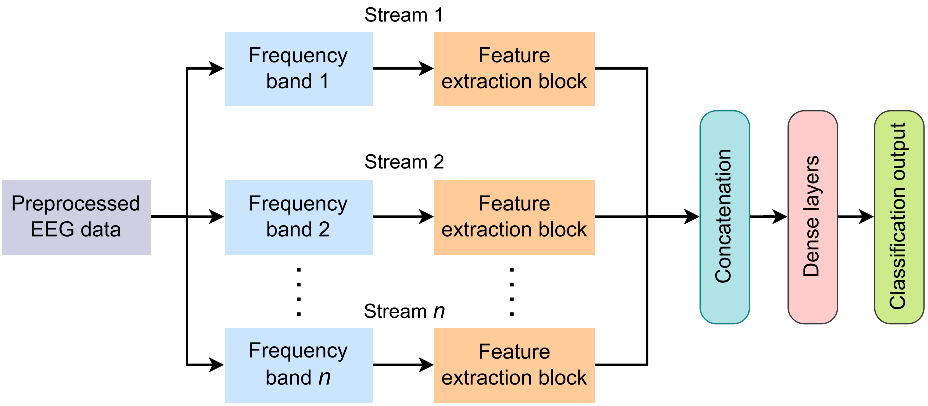


Fig. 1. Overview of the proposed classification framework.

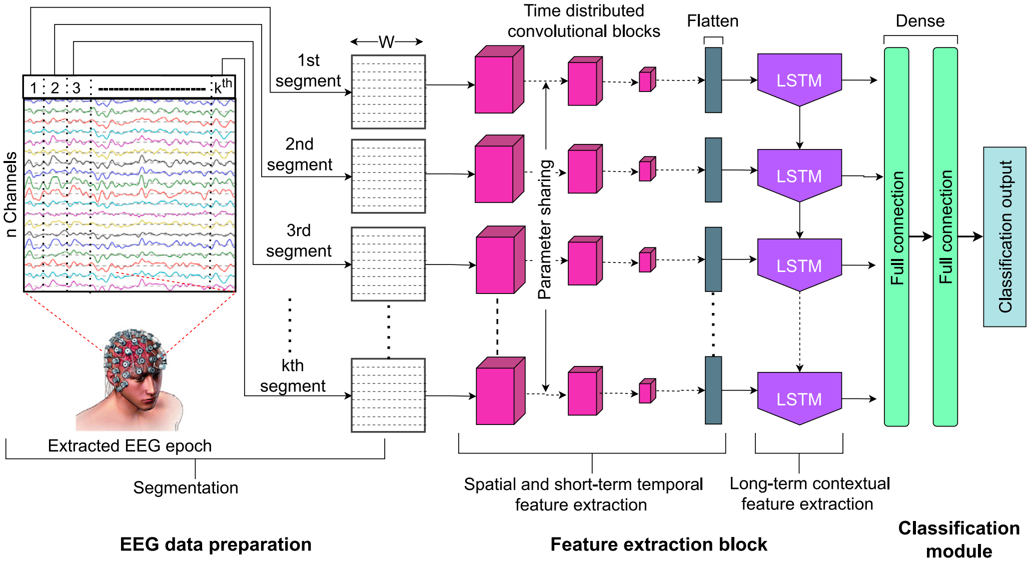


Fig. 2. The architecture of the feature extraction block. EEG epoch data is segmented into multiple fragments with a constant window length and then passed onto time-distributed convolution blocks. The time-sequenced features are then fed to LSTM. The classification module is appended to validate the efficacy of the feature extractor.

3.1. *The time-distributed CNN-LSTM feature extraction block*

Fig. 2 depicts the high-level architecture of the proposed time-distributed CNN-LSTM feature extraction block that is used in each stream of the multi-stream (MSSTNet) model. The spatio-temporal feature extraction is done through 3 steps: data segmentation, local spatial and temporal feature extraction, and long-term evolving feature mining. Each step is detailed next.

Let the total number of pre-processed EEG epochs be N. Each epoch consists of multivariate EEG data that is represented by $X_t \in \mathbb{R}^{n\times p}$, where, n = number of channels and p = number of time-points = sampling frequency × epoch length in seconds. The classification task is defined as obtaining a mapping of each of the N epochs to one of the two classes (MW and focus), that is, $g : X_t \rightarrow C_i; C_i \in \{C_1, C_2\}$, where C_i represents the class labels.

3.1.1. *Data segmentation*

To perform the convolution in a time-distributed manner, each X_t is divided into k number of smaller, non-overlapping fragments, each of width $W = \frac{p}{k}$. Each fragment is denoted as $X \in \mathbb{R}^{n\times W}$. The obtained sequential data segments are then fed to the convolutional blocks in a time-distributed manner for convolutional feature extraction.

3.1.2. *Local spatial and temporal feature extraction*

The spatial and short-temporal features are extracted through a series of convolutional and pooling layers. A convolutional layer contains a pool of filters that operate on the input via convolution. Given X be the input MW-EEG segment, characterized by n

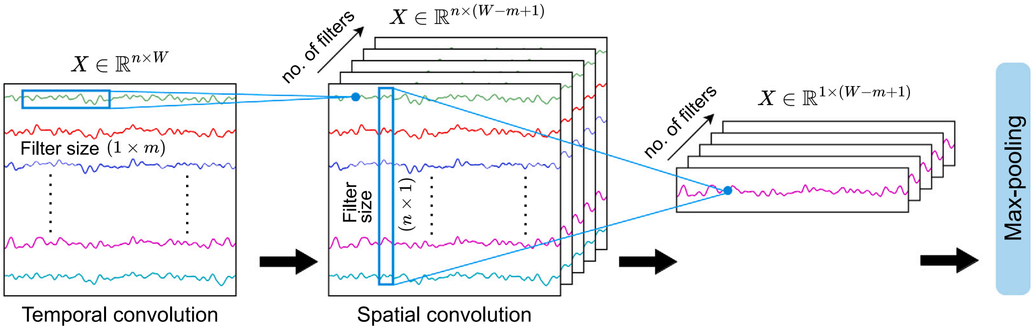

$X \in \mathbb{R}^{n\times W}$
Filter size $(1\times m)$
Temporal convolution
no. of filters
$X \in \mathbb{R}^{n\times (W-m+1)}$
Filter size
$(n\times 1)$
Spatial convolution
no. of filters
$X \in \mathbb{R}^{1\times (W-m+1)}$
Max-pooling

Fig. 3. Operation of the first convolutional block: short-term temporal and spatial convolution on an EEG segment using 1-D filters.

channels and W width, and f be the filter, the convolution operation is denoted by:

$$conv(X,f)_{x,y} = \sum_{i=1}^{W}\sum_{j=1}^{n} f_{i,j}X_{x+i-1,y+j-1} \tag{1}$$

where x and y denote the point of operation.

Now, a series of convolution operations coupled with a non-linear activation is applied to the input to generate the desired feature space. The convolution block is characterized by a set of q filters, f_q, which is trained iteratively. Since a series of convolution layers are used, the input is denoted by $X^{[0]}$, and the output of layer l will be denoted by $X^{[l]}$. These operations are represented by:

$$X^{[l]} = conv(X^{[l-1]}, f_q^{[l-1]}) \tag{2}$$

$$X^{[l-1]} = conv(X^{[l-2]}, f_q^{[l-2]}) \text{ and so on, till,} \tag{3}$$

$$X^{[1]} = conv(X^{[0]}, f_q^{[0]}) \tag{4}$$

If the activation of the lth layer is given by ϕ, then the convolution-activation operation can be written as,

$$X^{[l]} = \phi^{l}(conv(X^{[l-1]}, f_q^{[l-1]})) = \phi^{[l]}(\sum_{i=1}^{qW}\sum_{j=1}^{qC} f_{i,j}^{[l-1]}X_{x+i-1,y+j-1}^{[l-1]}) \tag{5}$$

To calculate the values of f_q for each layer l where $\forall l \in \{0,1,2,\dots,l-1\}$, the gradient is calculated w.r.t. each weight in f.

To capture the local spatial and temporal features, this work considers multiple standard convolutional blocks connected sequentially. Each block consists of four layers: one convolution layer, one batch normalization layer (for deep neural network stabilization), one max pooling layer, and one dropout layer (to reduce overfitting). For optimal performance, such blocks are connected sequentially to form a deep network. The number of convolutional blocks, filter sizes, and strides is obtained by trial-and-error during training.

In the first layer of the convolution block, a temporal filter of kernel size $1\times m$ (the value of m to be found out by hyperparameter tuning during training) is used, that performs a temporal convolution to find out the local temporal correlation in each channel's signal. Thereafter, a spatial filter of size $n\times 1$, where n =number of channels, is used to capture the spatial correlation between signals of multiple channels, spatially separated (Fig. 3). A max-pooling layer is then used to down-sample the feature map. Additionally, three more convolution-max pooling operations are performed to extract more complex/high-level features. Finally, a global average pooling layer is used to average all the features across the last set of filters, and the averaged feature map is fed to a 'flatten' layer, to generate a 1-D feature vector. The activation function used in each convolution layer is 'relu'.

The convolution features, extracted by this CNN structure, applied to each time-sequenced fragment (X), one fragment at a time, but with shared parameters, are fed into the LSTM model. To achieve this, a time-distributed wrapper is used for the CNN structure. Using this CNN model, features are extracted from input data. Finally, all extracted features are flattened to get a 1-D array of features. Hence, after convolutional feature extraction and flattening, a k number of 1-D feature vectors are obtained for k time-ordered segments, which are sequentially fed into the LSTM layer. Using these time-distributed features, the LSTM learns the temporal patterns across the k time segments and captures the long-term temporal evolution of features from EEG signals of prolonged MW episodes.

3.1.3. *Long-term contextual feature extraction*

The time evolution of the extracted convolutional features is captured by the memory blocks of the LSTM model. It uses the gating mechanism of LSTM on the generated feature space $X^{[l]}$ from the convolution block. The gating memory mechanism helps the architecture learn and store temporal information. Let g_t be the memory state at time t, the input be denoted by $\hat{X}_t$, and the

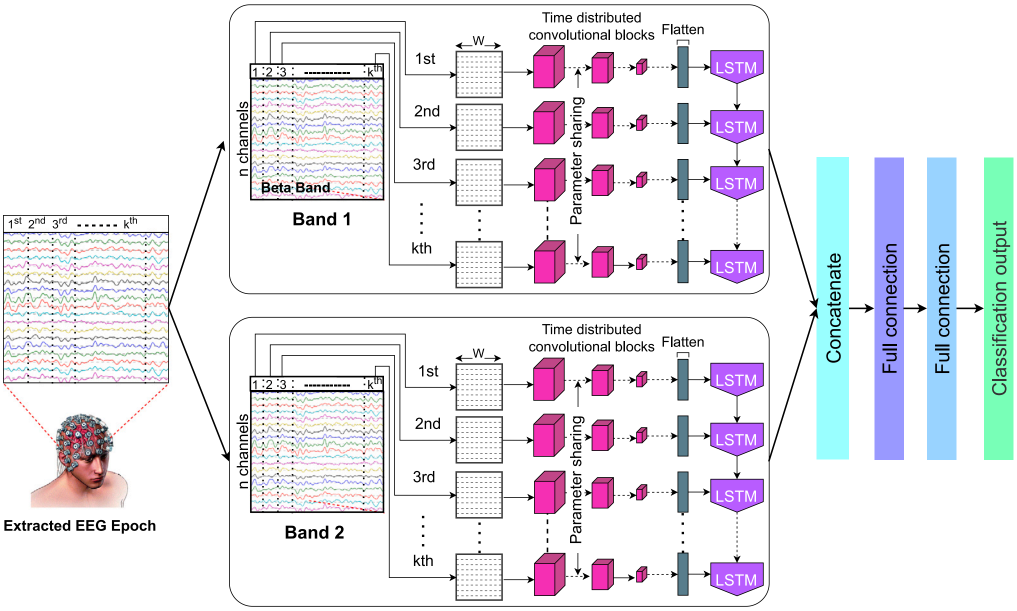


Fig. 4. High-level architecture of the MSSTNet model. The feature extraction block described previously (see Fig. 2) is used in the branches of the model.

output be y_t. Then,

$$y_t = \sigma(mem(\hat{X}_t)) \tag{6}$$

where σ is a non-linear activation. The *mem* function can be further broken down into,

$$g_t = \theta(U.\hat{X}_t + W.g_{t-1}) \tag{7}$$

$$y_t = \sigma(V.g_t) \tag{8}$$

In Eqs. (7) and (8), U, V, and W are the parameters, and σ and θ are the non-linear activations. Now, if for ith instance of data the memory operation is repeated, then

$$G_t^{(i)} = \theta(U, \hat{X}_t^i + W.G_t^{(i)}) \tag{9}$$

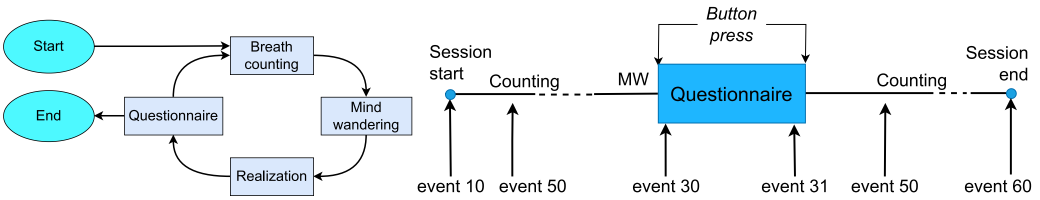
$$Y_t^{(i)} = \sigma(V.G_t^{(i)}) \tag{10}$$

Similar to the convolution block, the parameters of LSTM are also trainable using back-propagation. The memory state G_t is weighted by a parameter W so that the priority of the memory unit can also be learned by the model. With this time-distributed cascaded CNN-LSTM architecture, the spatio-temporal features are obtained from individual frequency bands.

3.2. The MSSTNet classification framework

The proposed MSSTNet model uses the above-described feature extractor in its streams to simultaneously extract features from multiple EEG frequency bands (see Fig. 4). In our study, we have created and evaluated models that contain 1–5 streams to extract features from multiple frequency bands. However, the subsequent experiments show marginal performance improvements for the models consisting of more than two streams. This is partly because of the overfitting issue that becomes prominent with the increasing number of hyperparameters for the smaller number of training data available. As the two-stream models have been proven to be more optimal (see Table 6) to detect MW in real-time scenarios, we perform extensive experiments on the two-stream MSSTNet model. Here, different combinations of pair-wise EEG bands (like Alpha + Beta, Beta + Delta, etc.) are simultaneously fed into the streams of MSSTNet for spatio-temporal feature extraction.

In this model, each MW or focus event is represented by two frequency-band features corresponding to band-1 (Alpha, say) and band-2 (Beta, say) data. The MSSTNet model is fed with this pair of bands' data for a particular event and produces a single classification output. The MSSTNet model consists of two streams: one for the first input (band-1 data) and another for the second input (band-2 data). Each stream of the MSSTNet model consists of feature extraction blocks (see Fig. 2) that calculate a feature set through convolutional and recurrent operations. Therefore, two sets of 1-D feature vectors are obtained for band-1 and band-2 EEG data in two branches. These two 1-D feature vectors are concatenated to form a larger (double-sized) 1-D feature vector. Now, this combined feature set is fed into several dense layers that produce the classification output.



(a) Flow of experimental process for each 20-minute session for each subject

(b) Button Press Event Sequence for each session, along with their corresponding event ID as represented in dataset

Fig. 5. The experimental flow in each session and the button press sequence [35].

4. Experiments and experimental results

4.1. The dataset

From previous research, it has been observed that a resting and relaxed state (awake-neutral state) is more prone to MW. In this case, a task involving low cognition, such as breath counting, can be used to represent the focused/alert state of mind [4]. Following these experimental criteria, a freely available dataset generated by the authors of [35] is used in the current study. The publicly accessible full dataset can be downloaded from https://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html.

Two healthy adults (1 male and 1 female) participated in the experiment, in which each of them appeared for 11 sessions. In each session, the subjects had to repeatedly count their breath cycle from 10 to 1 in reverse, sitting in a relaxed position and looking at a fixation cross on a computer screen in front of them. During counting, whenever they realized that an MW episode had happened and regained their meta-consciousness, they had to press a button to register the MW event. After the button was pressed, they had to fill out a questionnaire to report their self-analysis of the experience. When the questionnaire was completed, they repeated their breath-counting task after pressing a button indicating the start of a focus event. This chain of processes continued for 20 min each session. During the experiment, EEG data were recorded using a 64-electrode BIOSEMI device with a 1024 Hz sampling rate. The process flow and the corresponding event map are shown in Figs. 5(a) and 5(b) respectively.

4.2. Pre-processing

Pre-processing has been done using MNE-python (mne 0.20.7), an open-source Python library for EEG data processing. At first, a band-pass filter (0.1–40 Hz) is applied to raw EEG data to capture all main EEG frequency band information, which is termed ‘all-bands’ in the remaining literature. After the initial filtering, more band-pass filters are applied to get the five standard EEG bands’ data, considering 0.1–04 Hz as Delta, 04–08 Hz as Theta, 08–13 Hz as Alpha, 13–30 Hz as Beta, and 30–40 Hz as Gamma. Thereafter, the EEG recordings from $t = T - 10$ s to $t = T - 2$ s are considered MW data for each button press event indicating MW (assuming that the button press happens at $t = T$ s). The EEG signal of $T - 2$ to T second is assumed to be contaminated by muscle artifacts originated from button press activity and therefore eliminated. The 8-s recordings, immediately after the button press indicating the focus event, are considered focus data. The selection of event duration is inspired by the initial study of [4], in which a 10-s duration is considered. Further, we have removed 2-s data preceding each button press to avoid the muscle artifacts. This selection has been later experimentally confirmed by [14], where this particular selection provides higher classification performance compared to other options (e.g., 2 s, 4 s, and 6 s as duration). In total, 977 separate events, each lasting 8 s, are extracted from 22 sessions (combining both subjects); 472 of them are MW and the other 505 are focus events.

4.3. Performance metrics

Seven metrics are used to evaluate the model’s performance. These metrics are accuracy, sensitivity, specificity, precision, F1 score, Cohen’s kappa score (K), and AUC (area under the curve) score. A 10-fold cross-validation with an overall $90 : 10$ train-test split is done (with a randomized validation split of 0.2) for each experiment. The performance metrics averaged over 10 folds, along with standard error are computed.

4.4. Objectives of the experiments

- To find the efficiency of the feature extraction block with varying fragmentation window length and varying the input frequency bands’ data.
- To find the classification performance of the MSSTNet with varying pairs of frequency bands.
- To evaluate the impact of different hyperparameters and model components on the classification performance.
- To evaluate the model’s efficiency in handling subject-wise variability in mind-wandering EEG data.
- To validate the model’s performance by comparing it with baselines.
- To investigate the generalization capability of the model by evaluating it on an independent, related dataset.

Table 1
Hyperparameters of the feature extraction block for 2 s segmentation window.

No.	Layer name	# filters	Kernel size	Stride	Output size	# params
1	Convolution	10	1 × 10	1	4 × 64 × 2039 × 10	110
2	Convolution	10	64 × 1	1	4 × 1 × 2039 × 10	6410
3	Batch normalization	–	–	–	4 × 1 × 2039 × 10	40
4	Max-Pooling	–	1 × 2	2	4 × 1 × 1019 × 10	0
5	Dropout	–	–	–	4 × 1 × 1019 × 10	0
6	Convolution	10	1 × 10	1	4 × 1 × 1010 × 10	1010
7	Batch normalization	–	–	–	4 × 1 × 1010 × 10	40
8	Max - Pooling	–	1 × 2	2	4 × 1 × 505 × 10	0
9	Dropout	–	–	–	4 × 1 × 505 × 10	0
10	Convolution	10	1 × 10	1	4 × 1 × 496 × 10	1010
11	Batch normalization	–	–	–	4 × 1 × 496 × 10	40
12	Max-Pooling	–	1 × 4	4	4 × 1 × 124 × 10	0
13	Dropout	–	–	–	4 × 1 × 124 × 10	0
14	Convolution	10	1 × 10	1	4 × 1 × 115 × 10	1010
15	Batch normalization	–	–	–	4 × 1 × 115 × 10	40
16	Max-Pooling	–	1 × 3	3	4 × 1 × 38 × 10	0
17	Dropout	–	–	–	4 × 1 × 38 × 10	0
18	Global Average Pooling	–	–	–	4 × 10	0
19	Flatten	–	–	–	4 × 10	0
20	LSTM	–	–	–	80	29 120
	Total					38 830

Table 2
MSSTNet model, containing the feature extractor in Table 1 in it's branches.

Layer name	Output size	# params
Input Layer (band 1), Input layer (band 2)	4 × 64 × 2048 × 1, 4 × 64 × 2048 × 1	0
Feature extraction block	80, 80	38 830, 38 830
Concatenate	160	0
Dense	200	32 200
Dense	100	20 100
Dense (Output)	1	101
Total hyperparameters		1,30,061

4.5. Model implementation

4.5.1. Architecture and parameter settings of the feature extraction block

To identify an efficient model architecture and hyperparameter settings for the feature extraction block, a classification task is performed. Here, we utilize a single feature extraction block and append two dense layers with the feature extractor to perform the classification task. This single-stream network is referred as SSSTNet model in further discussion. As described in the model architecture, the LSTM model requires the input data in the form of time-sequenced subsets. In this study, we have considered multiple non-overlapping window sizes, specifically 0.5 s, 1 s, 2 s, and 4 s We observe the performance for each of the six frequency bands and these window sizes individually. The model performances of 0.5 s and 4 s are observed to be significantly lower (highest average detection accuracy of 85.78% and 84.36%, respectively) than that of 1 s and 2 s window sizes. Therefore, we have not further analyzed the performances of these two window sizes (refer to Section 5.2 for further discussion). Only the performances of 1-s and 2-s window sizes are shown in the results. The optimal hyperparameter setting for the feature extractor block is obtained through a grid-search that is done by searching over the following parameter values: no. of convolution layers: $[3, 4, 5, 6]$, no. of filters in each convolution layer: $[10, 20, 32, 64]$, kernel sizes except the spatial convolution layer: $[(1, 10), (1, 20)]$, strides (convolution filters): $[1, 2]$, strides (max-pooling): $[2, 3, 4]$, no of units in dense layers: $[25, 50, 75, 100, 125, 150, 200, 250]$, no of units in LSTM layer: $[60, 80, 100]$, dropout rate: $[0.2, 0.3, 0.5]$, and learning rate: $[0.001, 0.0001, 0.00001]$. A batch size of 32 is considered for model training, and training has been done for 150 epochs. A 'sigmoid' activation function is used in the last dense layer. An 'Adam' optimizer and 'binary cross-entropy' loss function are considered for training. The optimal learning rate for the optimizer and the dropout rate are 10^{-4} and 0.2, respectively. The optimal hyperparameters of the feature extractor are shown in Table 1. The SSSTNet model further consists of two hidden dense layers with numbers of nodes 200 and 100, respectively, and another dense layer acting as the output layer.

4.5.2. Implementation of MSSTNet classification framework

The MSSTNet model extracts features from a pair of MW-EEG frequency bands' data simultaneously. Here, we have utilized the feature extractor block, shown in Table 1, in each branch of the multi-stream model. The full MSSTNet model architecture is shown in 2. The codes of the proposed models can be found in https://github.com/subhra101/MSSTNet.

Table 3
The classification performance of the SSSTNet model for individual frequency bands. Among different bands', the Beta band achieves significantly higher accuracy.

Band	Accuracy↑	Recall↑	Specificity↑	Precision↑	F1 Score↑	K↑	AUC↑
1 s window							
δ	66.23 ± 0.41	57.80 ± 0.97	74.07 ± 0.47	67.64 ± 0.37	61.63 ± 0.83	0.32 ± 0.02	0.73 ± 0.01
θ	72.58 ± 0.71	62.98 ± 0.32	80.64 ± 0.26	75.81 ± 0.88	68.82 ± 0.65	0.42 ± 0.03	0.83 ± 0.01
α	82.79 ± 0.46	74.09 ± 0.90	90.89 ± 0.51	87.58 ± 0.62	79.90 ± 0.88	0.63 ± 0.02	0.91 ± 0.01
β	89.26 ± 0.50	87.69 ± 1.05	90.72 ± 0.64	89.90 ± 0.20	88.49 ± 0.71	0.75 ± 0.01	0.96 ± 0.01
γ	83.31 ± 0.37	77.90 ± 0.29	88.31 ± 1.13	85.72 ± 0.58	81.26 ± 1.50	0.65 ± 0.03	0.93 ± 0.03
All	89.47 ± 0.36	85.18 ± 0.48	93.53 ± 0.41	91.27 ± 0.51	88.15 ± 0.44	0.75 ± 0.02	0.97 ± 0.01
2 s window							
δ	67.98 ± 0.57	55.67 ± 0.51	79.50 ± 0.55	70.93 ± 0.59	61.90 ± 0.05	0.32 ± 0.03	0.74 ± 0.01
θ	73.91 ± 0.83	66.96 ± 0.52	80.44 ± 0.75	75.71 ± 0.18	70.68 ± 0.65	0.44 ± 0.03	0.80 ± 0.01
α	84.71 ± 0.39	74.38 ± 0.71	**94.35 ± 0.72**	**91.67 ± 0.59**	81.53 ± 0.98	0.66 ± 0.02	0.94 ± 0.01
β	90.59 ± 0.38	90.87 ± 0.86	90.34 ± 0.78	89.88 ± 0.61	90.28 ± 0.44	0.77 ± 0.01	0.97 ± 0.01
γ	86.69 ± 0.31	80.90 ± 0.59	92.13 ± 0.40	90.90 ± 1.06	84.97 ± 0.87	0.69 ± 0.01	0.95 ± 0.01
All	**91.34 ± 0.34**	**94.21 ± 0.63**	88.67 ± 0.72	88.81 ± 0.48	**91.37 ± 0.85**	**0.78 ± 0.01**	**0.98 ± 0.01**

Table 4
Comparison of 10-fold cross-validated classification performance using different pairs of EEG frequency bands, considering a 2-s time sequence window for LSTM.

Bands	Accuracy↑	Sensitivity↑	Specificity↑	Precision↑	F1 Score↑	K↑	AUC↑
$\alpha + \beta$	**95.07 ± 0.86**	**94.28 ± 0.90**	96.16 ± 0.72	95.17 ± 1.22	**94.50 ± 1.35**	**0.85 ± 0.03**	**0.99 ± 0.01**
$\alpha + \gamma$	93.03 ± 1.46	90.14 ± 1.58	95.74 ± 0.85	95.94 ± 0.80	92.67 ± 2.16	0.83 ± 0.02	0.97 ± 0.02
$\alpha + \theta$	85.93 ± 2.26	81.11 ± 2.58	90.52 ± 1.90	88.02 ± 1.26	84.10 ± 1.72	0.66 ± 0.03	0.92 ± 0.01
$\beta + \gamma$	**94.75 ± 1.20**	93.21 ± 1.48	94.97 ± 1.16	93.75 ± 1.80	93.38 ± 0.80	0.84 ± 0.03	0.98 ± 0.01
$\beta + \theta$	**93.22 ± 1.48**	88.15 ± 1.56	**98.39 ± 0.42**	**97.63 ± 0.78**	92.46 ± 1.20	0.80 ± 0.02	0.99 ± 0.01
$\theta + \gamma$	88.45 ± 2.56	88.67 ± 2.78	87.96 ± 1.80	88.96 ± 1.54	88.77 ± 1.68	0.72 ± 0.03	0.96 ± 0.02

4.6. Results

4.6.1. The efficacy of the proposed feature extraction block

From the obtained results of the SSSTNet model's performance tabulated in Table 3, several observations can be made. Firstly, most classification metrics are improved when a 2-s non-overlapping window length is considered for creating sequential input data for the LSTM model compared to a 1-s window length. So we utilize a 2-s fragmentation window length in the MSSTNet model. Secondly, the performance of the Delta band is less impressive, and therefore the Delta band is excluded from further analysis in the MSSTNet model. Thirdly, the detection accuracy (91.34%) considering all bands' data (0.1–40 Hz) is the highest. Among different frequency bands, the Beta band provides the highest classification accuracy (90.59%), outperforming the Alpha (84.71%) and Gamma (86.69%) bands. Finally, for the current imbalanced dataset, Cohen's Kappa Score and AUC ROC score are also important. The highest Cohen's Kappa score is obtained in the case of all bands (0.78), followed by Beta (0.77), indicating good classifier performance across all training instances/classes. The highest AUC score is obtained in the case of all bands (0.98), followed by Beta (0.97).

4.6.2. Classification performance of the MSSTNet model

Table 4 shows the performance of the MSSTNet model. In terms of accuracy, the Beta-Alpha band pair as input outperforms all other pairs with a 95.07% accuracy. Gamma-Beta band pair and Beta-Theta band pair also exhibit promising performance, with an accuracy of 94.75% and 93.22%, respectively. Additionally, the Beta-Alpha band pair shows good performance in terms of other metrics, with the highest sensitivity (94.28%), F1 score (94.50%), Cohen kappa's score (0.85), and AUC score (0.99). The highest specificity is obtained for the Theta-Beta pair (98.39%). Further, to investigate the performance enhancement in the MSSTNet model, a comparison of the accuracy of individual bands' in SSSTNet and pair-wise bands' EEG data in MSSTNet is drawn, which is depicted in Fig. 6. It is evident from Fig. 6 that the MSSTNet model provides enhanced performance across all the pairs compared to individual bands' performance.

Moreover, the Beta-Alpha, Alpha-Gamma, Beta-Gamma, and Beta-Theta band pairs in the multi-stream model outperform the highest accuracy (91.34%) of the single-stream model using all-bands (0.1–40 Hz) data. This particular result outlines the benefit of training separate feature extractors for individual EEG frequency bands rather than using a single extractor for the entire spectrum. Again, among the individual bands, the highest classification accuracy is obtained using the Beta band (90.59%, as shown in Table 3). Therefore, these results further highlight the importance of analysis of the Beta band alteration in MW characterization.

4.6.3. Ablation study of hyperparameters and model components of MSSTNet

To assure an optimal hyperparameter setting for MSSTNet, and to assess the influence of each model component on performance, an ablation study is conducted. Each branch of MSSTNet comprises five convolution layers and one LSTM layer, followed by two

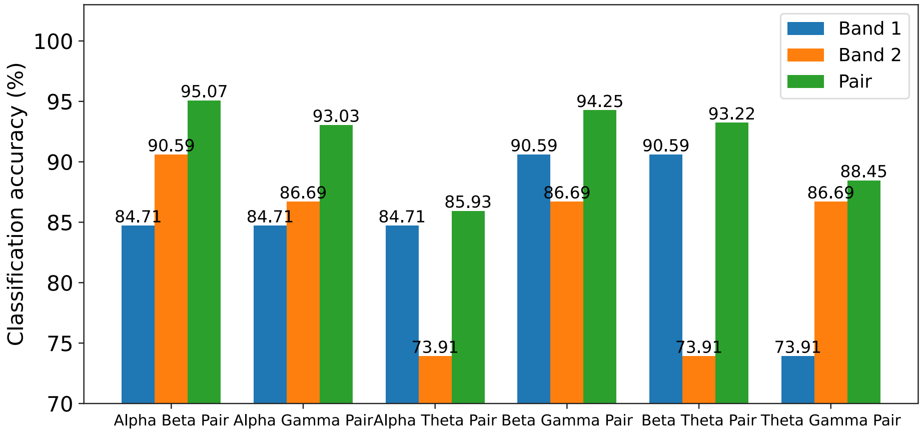


Fig. 6. Comparison of classification accuracy of pair-wise bands for the MSSTNet model versus the performance of individual frequency bands in single-stream model.

Table 5
The results of the ablation study of the MSSTNet model with Alpha-Beta pair data and a 2-s segmentation window. The highest-performing model is highlighted.

No. of convolution layers	Sequential layer name	No. of seq. units	No. of dense layers	Nodes in each dense layer	Accuracy
3	LSTM	80	2	(200, 100)	89.54
4	LSTM	80	2	(200, 100)	92.33
5	LSTM	**80**	**2**	**(200, 100)**	**95.07**
6	LSTM	80	2	(200, 100)	94.95
5	LSTM	60	2	(200, 100)	94.82
5	LSTM	100	2	(200, 100)	95.02
5	LSTM	80	1	100	92.67
5	LSTM	80	2	(50, 25)	93.10
5	LSTM	80	2	(100, 50)	94.45
5	LSTM	80	2	(150, 75)	94.72
5	LSTM	80	2	(250, 125)	95.02
5	GRU	60	2	(200, 100)	94.65
5	GRU	80	2	(200, 100)	94.80
5	GRU	100	2	(200, 100)	94.92

dense layers (see Fig. 4). To observe the effect of the model components and the hyperparameters, the numbers of convolution layers, LSTM units, and dense layers are varied. Moreover, we also utilize another variant of RNNs, namely GRUs, in place of the LSTM layer to explore their effectiveness in capturing sequential dependencies. While GRUs are computationally more efficient due to their simpler architecture, they may struggle to capture long-term dependencies as effectively as LSTMs. This analysis would help evaluate whether the reduced complexity of GRUs leads to any performance trade-offs. For these parameter variations, the accuracy of MSSTNet for the Alpha and Beta pair data with a 2-s segmentation size is reported (see Table 5). From Table 5, it is observed that both the LSTM and GRU-based model perform well, with the LSTM marginally outperforms the GRU. Further, a slight performance reduction is observed when the number of convolution layers becomes greater than five, mainly due to the increased number of hyperparameters, leading to overfitting. These results ensure the optimal hyperparameter setting of the MSSTNet model.

Increasing the number of streams in MSSTNet. The proposed classification framework consists of two branches that extract fine-grained features from a pair of frequency bands. From the implementation perspective, the number of streams in the model can easily be increased to capture features from more than two frequency bands, which may subsequently result in higher classification performance. To test this, an experiment is conducted to investigate the model's detection accuracy with varying numbers of streams, considering the highest-performing model in Table 5 as the baseline. For a particular number of streams, only the combination of the frequency bands with the highest accuracy is shown. Corresponding results are tabulated in Table 6. From Table 6, the accuracy is marginally increased with a 3-stream and 4-stream network, whereas a decline in performance is observed with a 5-stream network. This is due to the model overfitting caused by the increased number of hyperparameters in the model with a higher number of streams. The overfitting issue limits the expected increase in test accuracy due to the inclusion of features from multiple frequency bands. Moreover, the computational complexity and the training/testing time of the 3-stream and 4-stream networks are significantly higher than the 2-stream network, but the performances are only marginally better. To facilitate real-time and inexpensive MW detection, we consider the 2-stream network as the optimal model.

Table 6
The detection accuracy of the MSSTNet model with varying numbers of streams.

# Streams	1	2	3	4	5
Bands	All	$\alpha + \beta$	$\alpha + \beta + \gamma$	$\alpha + \beta + \gamma + \theta$	$\alpha + \beta + \gamma + \theta + \delta$
Accuracy	91.34	95.07	**95.36**	95.10	94.25

Table 7
Cross- and intra-subject performance of MSSTNet model using Alpha-Beta pair.

Mode	Training set	Testing set	Accuracy	Sensitivity	Specificity
Intra-subject	Sub-01	Sub-01	95.68	92.67	97.14
	Sub-02	Sub-02	93.27	93.79	92.86
Cross-subject	Sub-01	Sub-02	82.32	93.75	70.21
	Sub-02	Sub-01	83.94	82.81	81.78

4.6.4. Intra-subject and cross-subject performance analysis

To investigate the subject-wise variability of features extracted by MSSTNet, cross- and intra-subject classification is performed. The original experiment had two subjects: Sub-01 and Sub-02. Total reported events for both subjects differed, with 602 events for Sub-01 and 375 events for Sub-02. As DL model training is sensitive to the number of training instances, for comparison, an equal number of training instances for both subjects are considered. So, from the Sub-01 training set, 375 random instances are chosen while considering the whole Sub-02 data. For the intra-subject study, training and testing are done on the same subject, with a 90:10 train-test split. For cross-subject cases, training and testing are done on alternate subjects.

The intra-subject results (Table 7) indicate the robustness and generalization capability of the MSSTNet model, as for both subjects (mean = 94.48%) the performance is quite close to (93.27%, for Sub-01) or higher than (95.68%) the full dataset analysis (95.07% for total 977 instances considering both subjects and using Alpha-Beta band pair). The reduced size of training data may have a slight effect on model training. Cross-subject studies yield lower performance (82.32% and 83.94%, mean = 83.13%) than intra-subject studies, indicating the subject-wise variations of MW features.

4.6.5. Comparison with baseline models

To establish the improved detection capability of the proposed MW-detection scheme, the model performance is further compared with multiple self-defined and state-of-the-art baseline methods. The self-defined baselines utilize 52 statistical, time-, frequency-, and wavelet-domain features per channel as input and various ML models. Specifically, 4 standard ML models, namely, SVM, with ‘radial basis function’ as the kernel, Random Forest (RF), Adaptive Boosting (ADB), and Extreme Gradient Boosting (XGB), with their optimal hyperparameter settings, are utilized.

Along with the self-defined baselines, two generalized DL-based EEG decoders, namely EEGNet [10] and HTNet [11] are utilized for comparison. Further, another CNN-based model, state-of-the-art for driver’s attention detection, namely interpretable CNN (ICNN) [28], is also utilized. For the ICNN model, the code implementations shared by the authors are used, with one notable change. Following the author’s approach, we consider the pointwise convolution layer with 32 nodes and the subsequent depthwise convolution layer with 64 nodes. We also compare the proposed model with another CNN-based model [25], where a slightly different approach has been utilized. In this approach, STFT images obtained from raw EEG are fed to two pre-trained models (pre-trained on the Imagenet database), namely GoogLeNet and ResNet-34, for fine-tuning. For EEGNet, HTNet, and ICNN models, model architecture and hyperparameter settings can be found in their respective publications, whereas the pre-trained models (GoogLeNet and ResNet-34) are obtained from the *PyTorch* library. Further, motivated by the hybrid architecture design presented in [36], a sequential hybrid CNN-LSTM model is also tested. The hybrid CNN-LSTM (HCNN-LSTM) model follows the architecture and hyperparameters of the SSSTNet model, with one key difference. Like in [36], in this hybrid model, the entire EEG epoch data (of 8 s duration) is fed to CNN blocks, and the CNN-extracted features are fed to the LSTM model. In this model, no time-sequencing of raw EEG is done, and subsequently, it does not utilize the time-distributed convolutional blocks as in the MSSTNet model. Furthermore, inspired by the capability of Transformer-based models to decode very long time sequences efficiently, we also employ the TFormer model [32] for comparison. The time-domain EEG data, along with its frequency domain representation obtained by FFT, is used as input to the TFormer. To model the transformer blocks, the pre-norm Transformer architecture is utilized along with BatchNorm as the normalization technique. The hyperparameter setting can be found in the original study [32]. Finally, we utilize one RVFL-based driver detection model, namely, FGloWD-edRVFL [33]. As suggested in the original study, the ICNN-extracted features are utilized as FGloWD-edRVFL model input while following the model architecture as presented in the original study.

Note that all the comparison models considered here are single-streams by nature, that is, they accept only one frequency band data as input. To find a valid comparison, we apply a set of band-pass and band-stop filters to extract a combination of bands’ data. For example, to get the Alpha (08–13 Hz) and Gamma (30–40 Hz) pair, we apply a 08–40 Hz band pass filter, followed by a 13–30 Hz band stop filter on the raw EEG. Using these baseline models, the classification between MW and Focus is performed.

To establish the statistical superiority of the model’s performance, Wilcoxon’s rank-sum test-based method [37] is performed. More precisely, in this approach, different models are evaluated with the same training and test sets, and their performances are statistically compared. The statistical method essentially tests the hypothesis that the accuracy of one model is greater than the other or not, at a 5% level of significance, under similar experimental conditions. Following the 10-fold cross-validation scheme,

Table 8
Accuracy comparison between MSSTNet and baseline models. The architectures of the DL baseline models are detailed in their original studies.

	$\alpha + \beta$	$\alpha + \gamma$	$\alpha + \theta$	$\beta + \gamma$	$\beta + \theta$	$\theta + \gamma$	$p-$ values, model vs. MSSTNet
SVM	85.11	84.24	72.56	85.28	83.32	81.46	0.000079
RF	85.58	85.24	76.90	86.36	82.88	82.46	0.000079
ADB	83.68	84.72	73.85	84.28	80.56	79.80	0.000079
XGB	87.36	85.68	73.20	85.78	84.10	83.82	0.000079
EEGNet [10]	90.80	87.86	75.48	89.24	86.35	86.78	0.000440
HTNet [11]	91.20	86.25	77.36	89.56	85.88	87.84	0.000335
GoogLeNet [25]	88.30	86.68	78.90	78.35	88.80	86.36	0.000253
ResNet-34 [25]	88.56	85.48	79.50	80.32	87.76	84.18	0.000079
ICNN [28]	91.50	88.66	84.58	93.76	92.25	86.36	0.020625
HCNN-LSTM [36]	88.46	85.78	73.36	88.10	82.92	84.78	0.000079
TFormer [32]	94.26	**94.10**	81.60	91.84	89.25	86.34	0.048152
FGloWD-edRVFL [33]	92.25	89.34	85.70	94.20	92.82	87.68	0.041049
SSSTNet	92.56	90.48	84.24	91.68	91.20	85.36	0.003251
MSSTNet	**95.07**	93.03	**85.93**	**94.25**	**93.22**	**88.45**	–

Table 9
Comparison between MSSTNet and other reported MW detection models.

	Extracted features	Model	Testing scheme	Performance
[7]	P1, N1, P3 responses, power, and coherence in Theta and Alpha	SVM	Cross-subject	ACC: 67.5%
[38]	Minimum amplitude of N1 and maximum amplitude of P3 responses	SVM	Intra-subject Cross-subject	AUC: 0.715 AUC: 0.613
[2]	Log-normalized power of CSP components	SVM	Intra-subject	ACC: 84.00%
[14]	Feature extraction and classification using CNN		Mixed-subject Cross-subject	ACC: 91.78% ACC: 66.45%
[16]	Features extracted by ResNet-50 fed to RNN		Mixed-subject	ACC: 92.04%
[15]	Handcrafted feature, CNN		Cross-task	ACC: 0.519
Pro-posed	Multi-stream time-distributed CNN-LSTM model		Mixed-subject Intra-subject Cross-subject	ACC: 95.07% ACC: 94.48% ACC: 83.13%

the dataset is split into ten folds, 9 folds for training and the remaining 1 fold for testing. For each fold, the accuracy of the baselines (with the best performing frequency bands) and the MSSTNet model are evaluated. Then the accuracies for 10 folds for two models (for example, the SVM model with β and γ band's data versus MSSTNet with α and β bands' data) are compared using Wilcoxon's rank-sum test. The obtained $p-$ values are reported in Table 8. It is evident from Table 8 that the SSSTNet and MSSTNet models significantly outperform the handcrafted feature-based ML models for all the bands. Further, the performances of single-stream CNN-based and hybrid models (EEGNet, HTNet, and HCNN-LSTM) are significantly lower than the multistream model, indicating the efficacy of the fine-grained features extracted separately from individual bands. The attention-based TFormer model performs better than other CNN-based models, achieving the highest performance for the Alpha and Gamma pair. The ICNN-extracted feature-based FGloWD-edRVFL model also exhibits a fair performance for MW detection. Moreover, the combination of Beta and Alpha bands ($\beta + \alpha$) and Beta and Gamma bands ($\beta + \gamma$) consistently provides better discrimination compared to other bands' combinations for all classifiers, indicating the existence of the discriminatory MW features in these bands. Finally, considering the performances of different classifiers and the statistical testing results ($p-$ values < 0.05 for all models when compared with MSSTNet), the superior detection capability of the MSSTNet model is established.

Furthermore, along with high accuracy, in real-time, cost-efficient systems, lower computational complexity and fewer hyperparameters are crucial, as they lead to reduced training, testing times, and computational resources. The proposed MSSTNet model, with 1,30,061 hyperparameters, offers a more efficient architecture than GoogLeNet and ResNet-34, which have 5,602,979 and 21,286,211 hyperparameters [25], respectively. This is due to the lack of recurrent layers, essential for capturing the time-evolving patterns in EEG, leading to inferior performance than MSSTNet (see Table 8). EEGNet offers a more compact, time-efficient model with 10,322 hyperparameters (EEGNetV4, when utilized for the current task). The inferior performance of EEGNet compared to MSSTNet is due to the absence of recurrent layers and multiple streams in EEGNet, which limits its ability to extract bandwise fine-grained features. The use of self-attention mechanisms in transformer-based TFormer models often leads to higher computational demands and complexity, especially with longer inputs, as in the current MW detection task. So, the proposed MSSTNet model efficiently balances performance and computational efficiency, making it suitable for real-time, cost-efficient MW detection.

Comparison with the reported MW detection models. Given that the efficacy of any data-driven model fluctuates with the experiment and associated task, classification studies utilizing the same dataset are particularly significant for reliable comparison. The studies

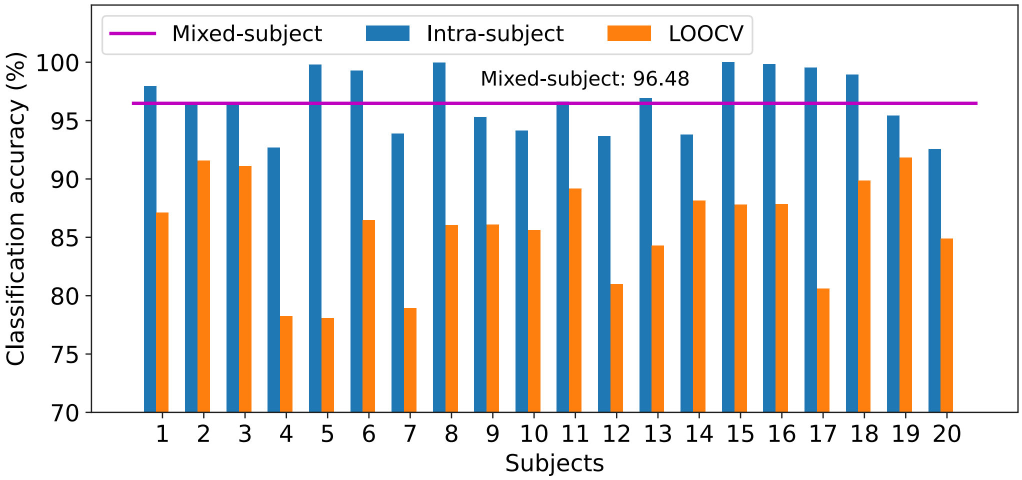


Fig. 7. The MW detection performance of the proposed model for breath focus meditation and instructed MW task [39] for 20 subjects. Superior performances using the MSSTNet model for mixed-, intra-, and LOOCV classification strategies are observed.

presented in [14,16] employ the same breath counting dataset as ours, but the other studies in Table 9 utilize their proprietary datasets. Table 9 presents the performance metrics of reported MW detection models. The variation in experimental paradigms and tasks between studies complicates the establishment of a valid comparison. Table 9 indicates that the reported MW detection performance is typically subpar and varies considerably between subjects. The CNN model in [14] demonstrates fair detection performance; nevertheless, its cross-subject accuracy was inadequate, constraining its generalization abilities. The MSSTNet model surpasses reported MW detection models in accuracy and generalization, which is established in both mixed- and cross-subject evaluations.

4.6.6. Generalizability of the proposed MW detection model

The performance of any DL model depends upon the particular dataset and the volume of available training data. Although the primary dataset used in the study has plenty of session-wise variability, it lacks subject-wise variability. Therefore, we further validate the MSSTNet's performance using another related, publicly available (https://zenodo.org/records/57911) dataset to test its generalization capability. We hereby consider the experiment performed in the Meditation Research Institute (MRI) in Rishikesh, India, where the brain signals of instructed mind wandering and breath focus meditation have been investigated. The EEG has been recorded for 20 min for each brain state from persons performing different types of meditation. We consider the EEG recordings from 20 meditators from the Isha Shoonya Yoga tradition for our analysis. Further details about the dataset can be found in the original study [39]. The 20-minute EEG recording is fragmented into 8-s segments, with 6-s overlap between consecutive segments. Thus, 597 segments are obtained for each person's brain state. Using this training data, mixed- and intra-subject classification is performed using the 10-fold cross validation. Further, to test the inter-subject performance similar to Table 7, a leave-one-out cross validation-based classification (LOOCV) is performed, where 19 subjects' data is utilized for training and the remaining 1 subject's data is utilized for testing. For this, we have utilized the best detection model for our primary dataset, that is MSSTNet, using $\alpha + \beta$ band's data with a 2-s window for LSTM, while other model hyperparameters are kept the same as previous (see Tables 1 and 2).

The classification performance of the three training-testing strategies is shown in Fig. 7. The model exhibits a 96.48% ±1.86 accuracy for the mixed-subject classification, whereas an average accuracy of 96.65% ±2.57 and 85.73% ±4.24 (averaged over all subjects) is achieved for intra-subject and LOOCV classification schemes. Observed intra- and mixed-subject performance of the proposed MSSTNet is significantly better than the detection model presented in [40], where a set of time domain features are utilized for a ML-based classification for the same dataset, and a mixed-subject accuracy of 94.06% was obtained. The obtained performance also highlights the effect of the instructed MW state (thinking about autobiographical memories) on the Beta and Alpha bands' data, compared to the breath-focus state. Further, the Beta and Gamma bands' data also exhibits better mixed-subject classification accuracy (95.90%), indicating the difference in Beta and Gamma waves during meditation and instructed MW states [39]. These results highlight the generalizability and robustness of the proposed classification framework to detect MW across diverse experimental conditions and subjects.

5. Discussion

In this section, the different factors affecting the detection model's performance are discussed. Moreover, the impact of MW-onset on the oscillatory activity of each frequency band are analyzed.

5.1. Impact of the segmentation schemes on model performance

The segmentation strategy used to produce time-sequenced features for the LSTM model substantially affects the performance. This study found that a 2-s segmentation window outperforms other choices, such as windows of 1, 0.5, and 4-s duration. For segmentation window widths of 0.5 s or less, the testing accuracy is significantly lower. As the window length lowers, the number of EEG segments rises, and LSTM and dense layer hyperparameters grow dramatically. The models overfit, resulting in poor testing accuracy. However, bigger window sizes produce fewer EEG segments, which cannot provide the contextual information to LSTM models.

5.2. Factors affecting the model training

The classification capability of any deep neural network model depends on the optimal training, such that the model does not underfit or overfit. Therefore, all the model's hyperparameters are delicately tuned to achieve optimal performance. A z-score normalization is performed on the input EEG data to achieve faster model convergence and stable training. Further, the hidden layer activations are normalized using batch normalization after each convolution for faster model convergence. The drop-out layers and an early-stopping scheme (by monitoring the validation loss) ensure minimal overfitting.

5.3. Impact of MW-onset on different frequency bands

The impact of MW-onset on individual frequency bands can be explained by the SSSTNet model's performances shown in Table 3. The all-band's accuracy is found to be the highest (91.34%). The feature set extracted from all-band's data is a combination of all features that can be extracted using individual bands. Therefore, this collection of features encompasses most of the significant alterations across all bands and thus results in the highest accuracy. Further, the accuracy using Beta band's (90.59%) data significantly outperforms other bands, followed by Gamma and Alpha bands. MW-induced alteration in Alpha rhythm has been investigated in multiple recent studies. Some of these studies [3,6,21] report an increased Alpha power during the MW event, compared to focus. They conclude that a change in the Alpha band signal can be a better indicator of MW onset compared to a change in the Theta band. The effect on other bands was not largely explored previously. In our study, Alpha band accuracy (84.71%) is much higher than that of the Theta band (73.91%), indicating the Alpha band alterations are more prominent compared to Theta alterations. Therefore, our DL-based classification results are consistent with these previous results, where EEG signal analysis was done in terms of band power or absolute log power. Moreover, our results reveal the importance of the Beta band alterations due to MW-onset. Beta band classification accuracy is found to be higher than Alpha band accuracy. A low-amplitude beta wave is typically connected with low-stress active thinking. As the focus state requires concentration and MW indicates a loss of concentration, alterations in the Beta band are expected, and this could be the reason behind the high performance of the Beta band.

6. Conclusions

This study proposes an automated framework to extract discriminatory spatio-temporal features from EEG signals to detect mind wandering. The proposed time-distributed CNN-LSTM feature extraction block can extract spatial, short-term, and long-term temporal features from MW-EEG signals. Using this feature extraction block in each stream, a multi-stream Deep Learning model is proposed that is specialized to extract fine-grained features from individual EEG frequency bands. Compared to several state-of-the-art baseline models, the proposed model has achieved superior performance in mixed-, intra-, and inter-subject classification. Our analysis also reveals that the Beta and Alpha bands undergo the most prominent alterations due to MW onset. The proposed framework can be implemented in real-time settings to detect mind wandering. Thus, the detection model has a huge scope for practical application, especially during low-cognition tasks such as freeway driving, online lecture listening, reading, etc., which may be investigated in the future. The model's efficacy can be validated for other EEG classifications.

CRediT authorship contribution statement

Subrata Pain: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Subhrasankar Chatterjee:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Monalisa Sarma:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Debasis Samanta:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

This study was not funded by any funding agency.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is publicly available online. The link to the data and codes developed have been provided in the manuscript.

References

[1] Killingsworth MA, Gilbert DT. A wandering mind is an unhappy mind. Science 2010;330(6006). 932–932.

[2] Dhindsa K, Acai A, Wagner N, Bosynak D, Kelly S, Bhandari M, Petrisor B, Sonnadara RR. Individualized pattern recognition for detecting mind wandering from EEG during live lectures. PLoS One 2019;14(9):e0222276.

[3] Baldwin CL, Roberts DM, Barragan D, Lee JD, Lerner N, Higgins JS. Detecting and quantifying mind wandering during simulated driving. Front Hum Neurosci 2017;11:406.

[4] Braboszcz C, Delorme A. Lost in thoughts: neural markers of low alertness during mind wandering. Neuroimage 2011;54(4):3040–7.

[5] Atchley R, Klee D, Oken B. EEG frequency changes prior to making errors in an easy stroop task. Front Hum Neurosci 2017;11:521.

[6] Arnau S, Löffler C, Rummel J, Hagemann D, Wascher E, Schubert A-L. Inter-trial alpha power indicates mind wandering. Psychophysiology 2020;57(6):e13581.

[7] Jin CY, Borst JP, van Vugt MK. Predicting task-general mind-wandering with EEG. Cogn Affect Behav Neurosci 2019;19(4):1059–73.

[8] Jin CY, Borst JP, van Vugt MK. Distinguishing vigilance decrement and low task demands from mind-wandering: A machine learning analysis of EEG. Eur J Neurosci 2020;52(9):4147–64.

[9] Ke J, Zhang M, Luo X, Chen J. Monitoring distraction of construction workers caused by noise using a wearable electroencephalography (EEG) device. Autom Constr 2021;125(February):103598.

[10] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. EEGnet: a compact convolutional neural network for EEG-based brain–computer interfaces. J Neural Eng 2018;15(5):056013.

[11] Peterson SM, Steine-Hanson Z, Davis N, Rao RP, Brunton BW. Generalized neural decoders for transfer learning across participants and recording modalities. J Neural Eng 2021;18(2):026014.

[12] Algarni M, Saeed F, Al-Hadhrami T, Ghabban F, Al-Sarem M. Deep learning-based approach for emotion recognition using electroencephalography (EEG) signals using Bi-directional long short-term memory (Bi-LSTM). Sensors 2022;22(8):2976.

[13] Chang Y, Stevenson C, Chen I-C, Lin D-S, Ko L-W. Neurological state changes indicative of ADHD in children learned via EEG-based LSTM networks. J Neural Eng 2022;19(1):016021.

[14] Hosseini S, Guo X. Deep convolutional neural network for automated detection of mind wandering using EEG signals. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. 2019, p. 314–9.

[15] Jin CY, Borst JP, van Vugt MK. Decoding study-independent mind-wandering from EEG using convolutional neural networks. J Neural Eng 2023;20(2):026024.

[16] Zhu L, Zhu F, Price J. TopographyNET: a deep learning model for EEG-based mind wandering detection. In: Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics. 2022, p. 1–10.

[17] Henriquez Chaparro RA. Behavioral and neural correlates of spontaneous attentional decoupling: towards an understanding of mind wandering (Ph.D. thesis), Paris 6; 2015.

[18] Gouraud J, Delorme A, Berberian B. Mind wandering influences EEG signal in complex multimodal environments. Front Neuroergonom 2021;2:625343.

[19] Aliyu I, Lim CG. Selection of optimal wavelet features for epileptic EEG signal classification with LSTM. Neural Comput Appl 2021;1–21.

[20] Singh K, Malhotra J. Two-layer LSTM network-based prediction of epileptic seizures using EEG spectral features. Complex Intell Syst 2022;1–14.

[21] Compton RJ, Gearinger D, Wild H. The wandering mind oscillates: EEG alpha power is enhanced during moments of mind-wandering. Cogn Affect Behav Neurosci 2019;19(5):1184–91.

[22] Datta D, David PE, Mittal D, Jain A. Neural machine translation using recurrent neural network. Int J Eng Adv Technol 2020;9(4):1395–400.

[23] Maiorana E. Deep learning for EEG-based biometric recognition. Neurocomputing 2020;410:374–86.

[24] Khalil M, Adib A, et al. An end-to-end multi-level wavelet convolutional neural networks for heart diseases diagnosis. Neurocomputing 2020;417:187–201.

[25] Khan SA, Chaudary E, Mumtaz W. EEG-ConvNet: Convolutional networks for EEG-based subject-dependent emotion recognition. Comput Electr Eng 2024;116:109178.

[26] Li R, Gao R, Suganthan PN. A decomposition-based hybrid ensemble CNN framework for driver fatigue recognition. Inform Sci 2023;624:833–48.

[27] Xin Q, Hu S, Liu S, Zhao L, Zhang Y-D. An attention-based wavelet convolution neural network for epilepsy EEG classification. IEEE Trans Neural Syst Rehabil Eng 2022;30:957–66.

[28] Cui J, Lan Z, Sourina O, Müller-Wittig W. EEG-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network. IEEE Trans Neural Netw Learn Syst 2022;34(10):7921–33.

[29] Toraman S. Automatic recognition of preictal and interictal EEG signals using 1D-capsule networks. Comput Electr Eng 2021;91:107033.

[30] Moctezuma LA, Suzuki Y, Furuki J, Molinas M, Abe T. GRU-powered sleep stage classification with permutation-based EEG channel selection. Sci Rep 2024;14(1):17952.

[31] Supakar R, Satvaya P, Chakrabarti P. A deep learning based model using RNN-LSTM for the detection of Schizophrenia from EEG data. Comput Biol Med 2022;151:106225.

[32] Li R, Hu M, Gao R, Wang L, Suganthan P, Sourina O. TFormer: A time–frequency transformer with batch normalization for driver fatigue recognition. Adv Eng Inform 2024;62:102575.

[33] Li R, Gao R, Yuan L, Suganthan PN, Wang L, Sourina O. An enhanced ensemble deep random vector functional link network for driver fatigue recognition. Eng Appl Artif Intell 2023;123:106237.

[34] Li R, Gao R, Suganthan PN, Cui J, Sourina O, Wang L. A spectral-ensemble deep random vector functional link network for passive brain–computer interface. Expert Syst Appl 2023;227:120279.

[35] Grandchamp R, Braboszcz C, Delorme A. Oculometric variations during mind wandering. Front Psychol 2014;5:31.

[36] Pandey SK, Janghel RR, Mishra PK, Ahirwal MK. Automated epilepsy seizure detection from EEG signal based on hybrid CNN and LSTM model. Signal Image Video Process 2023;17(4):1113–22.

[37] Karnati M, Sahu G, Gupta A, Seal A, Krejcar O. A pyramidal spatial-based feature attention network for schizophrenia detection using electroencephalography signals. IEEE Trans Cogn Dev Syst 2023.

[38] Dong HW, Mills C, Knight RT, Kam JW. Detection of mind wandering using EEG: Within and across individuals. PLoS ONE 2021;16(5 May 2021):1–18.

[39] Braboszcz C, Cahn BR, Levy J, Fernandez M, Delorme A. Increased gamma brainwave amplitude compared to control in three different meditation traditions. PLoS One 2017;12(1):e0170647.

[40] Kaur K, Khandnor P. Temporal-domain analysis of meditation and mind-wandering EEG signals for different meditation traditions. In: 2023 7th international conference on computer applications in electrical engineering-recent advances. CERA, IEEE; 2023, p. 1–6.