frontiers in Computer Science

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fcomp.2022.780580

Multimodal EEG and Eye Tracking Feature Fusion Approaches for Attention Classification in Hybrid BCIs

Lisa-Marie Vortmann[1]*, *Simon Ceh*[2] *and Felix Putze*[1]

[1] *Cognitive Systems Lab, Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany,*
[2] *Department of Differential Psychology, University of Graz, Graz, Austria*

Often, various modalities capture distinct aspects of particular mental states or activities. While machine learning algorithms can reliably predict numerous aspects of human cognition and behavior using a single modality, they can benefit from the combination of multiple modalities. This is why hybrid BCIs are gaining popularity. However, it is not always straightforward to combine features from a multimodal dataset. Along with the method for generating the features, one must decide when the modalities should be combined during the classification process. We compare unimodal EEG and eye tracking classification of internally and externally directed attention to multimodal approaches for early, middle, and late fusion in this study. On a binary dataset with a chance level of 0.5, late fusion of the data achieves the highest classification accuracy of 0.609–0.675 (95%-confidence interval). In general, the results indicate that for these modalities, middle or late fusion approaches are better suited than early fusion approaches. Additional validation of the observed trend will require the use of additional datasets, alternative feature generation mechanisms, decision rules, and neural network designs. We conclude with a set of premises that need to be considered when deciding on a multimodal attentional state classification approach.

Keywords: feature fusion, convolutional neural networks, attention, eye tracking, EEG, Markov Transition Fields, Gramian Angular Fields

OPEN ACCESS

Edited by:
Laura M. Ferrari,
Université Côte d'Azur, France

Reviewed by:
Maryam S. Mirian,
University of British Columbia, Canada
Siyuan Chen,
University of New South Wales, Australia

***Correspondence:**
Lisa-Marie Vortmann
vortmann@uni-bremen.de

Specialty section:
This article was submitted to Mobile and Ubiquitous Computing, a section of the journal Frontiers in Computer Science

Received: *21 September 2021*
Accepted: *21 February 2022*
Published: *21 March 2022*

Citation:
Vortmann L-M, Ceh S and Putze F (2022) Multimodal EEG and Eye Tracking Feature Fusion Approaches for Attention Classification in Hybrid BCIs. Front. Comput. Sci. 4:780580. doi: 10.3389/fcomp.2022.780580

1. INTRODUCTION

Human-machine interaction is becoming increasingly ubiquitous. In our daily lives, we want to seamlessly incorporate technology and thus rely on usability. By integrating implicit input mechanisms, the synergy between users and machines is further enhanced: These enable a system to infer information about the user without the user taking any explicit action, such as pressing a button or speaking a command, and modify their behavior accordingly.

One way of implementing implicit input mechanisms is *via* biosignal-based recognition of cognitive states. Biosignal-based recognition of cognitive states or activities in humans is a broad research field because of the manifold options for input signals, classification algorithms, and possible applications. For instance, a Brain-Computer Interface (BCI) can predict a user's attentional state from electroencephalographic (EEG) data and adapt the system's behavior using machine learning (Vortmann and Putze, 2020). Certain modalities are more suited to certain

applications and scopes than others, but for the majority of applications, more than one possible input signal can be considered. For instance, brain activity can be supported by eye gaze behavior. Such systems are commonly referred to as hybrid BCIs (Kim et al., 2015).

The fundamental premise of such multimodal approaches in the context of BCI machine learning is that the two modalities may capture distinct aspects of the user state and thus complement one another. While using a single modality can result in reliable classification accuracy, combining two or more modalities can enhance the system's recognition power and robustness, thereby improving its overall performance. D'Mello and Kory (2012) demonstrated in a review of 30 studies that multimodal classification yielded on average 8.12% improvement over the unimodal classifiers. Possible aims of the combination are to correct for temporally noisy data, resolve ambiguity, or the exploitation of correlations (Baltrušaitis et al., 2018).

In this work, we want to systematically explore the combination of EEG and eye tracking data for the classification of internally and externally directed attention. The result of such a classification could be used in a BCI to adapt the system to the user state.

1.1. Multimodal Feature Fusion

Biosignal data is heterogeneous in nature due to its inherent properties and recording mechanisms. For example, brain activity can be recorded using an EEG, which measures electrophysiological changes on the scalp and is usually recorded in microvolt, whereas eye gaze behavior is recorded by eye tracking devices that measure pupil dilation and infer gaze coordinates. During unimodal approaches, the feature extraction is either explicitly designed to generate meaningful features from the data, or the classification process implicitly learns to extract modality- and task-specific features (Kim et al., 2020). A combination of several modalities for the classification process is therefore not trivial.

The first opportunity to merge modalities is before the beginning of the classification process. Such **early fusion** approaches combine the biosignals on a feature level (Cheng et al., 2020). The joint representation of previously extracted meaningful features or preprocessed raw data presupposes that all modalities can be aligned properly for classification. This approach allows for the learning of cross-modal correlations during the classification process, but requires concatenation of the inputs and limits the extraction of modality-specific features.

Oppositely, **late fusion** approaches merge the modalities at the end of the classification process. The inputs are separately processed in individually tailored steps, typically until the prediction of individual labels. The fusion happens on the decision level based on the multiple predictions (Cheng et al., 2020). In Mangai et al. (2010), this was discussed as classifier combination because several classifiers are trained individually per modality before the results of the classifiers are combined (or one classifier is selected as overall output). The authors suggested different approaches how to choose the classifier combination, based on the available individual output formats per modality classifier. For instance, if each classifier predicts only a class label, an odd number of classifiers should be chosen to allow for (weighted) majority votes for the final output. In other cases, the classifiers could produce vectors in which the values represent the support for each label. Such certainty evaluations per modality classifier allow for a more sophisticated assessment of the final combined multimodal output. A decision rule has to define how the individual predictions are combined for the final prediction. This rules can either be set or learned using machine learning. The setting of a decision rule requires good *a priori* knowledge on the expected results, while machine learning based late fusion requires a large amount of data to enable the training of such decision rule. Especially regarding the proposed attention classification biosignal data, such large datasets are often not available and rule-based late fusion approaches should be favored. An apparent advantage of late fusion is the power of a tailored classification processes, whereas the shortcoming lies in the exploitation of modality correlations (Polikar, 2012).

One can also steer a middle course in fusing the modalities in the middle of the classification process. The idea of **middle fusion** (or halfway fusion) approaches is to first process the modalities individually but merge intermediate results as soon as possible, followed by further classification steps. In terms of neural networks, the first layers process the distinct inputs simultaneously before concatenating the layers' outputs for the following shared layers. The advantage of this fusion approach is that the modalities could first be processed tailored to their individual properties before exploiting the cross-correlations and arriving at a joint prediction.

1.2. EEG and Eye Tracking Based Mental State Detection

Hybrid BCIs have been used to detect a variety of mental states by analyzing eye movement patterns rather than relying on the user's explicit gaze behavior for direction control or target selection. As mentioned before, MI is a suitable use case for BCIs in general. Dong et al. (2015) used the natural gaze behavior of the participants to smooth the noisy predictions that resulted only from EEG motor imagery tasks. Cheng et al. (2020) explicitly compared late and early fusion of the multimodal features for their MI task. For the feature level fusion, they remarked that EEG and eye tracking data are so dissimilar, fusing them is not trivial and requires several preprocessing steps. For the decision level fusion, they used a decision rule based on the D-S evidence theory (Zhang et al., 2018). They found that feature fusion outperforms single modalities and that late fusion outperforms early fusion of eye tracking and EEG data.

In Guo et al. (2019), the authors investigate emotion recognition using a multimodal approach. They combine eye tracking and EEG data and classify the input after an early fusion using a deep neural network model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. For the early fusion of the modalities, they apply a Bimodal Deep AutoEncoder (BDAE) that extracts a high-level representation of features. This approach was first presented in Liu et al. (2016). Another early fusion approach for emotion recognition was examined in Lu et al. (2015). They fused 33

different features from eye movement data with 62 channel EEG signals and achieved 87.59% accuracy in classifying three emotions. Zheng et al. (2014) combined EEG signals and pupil dilation either in an early fusion approach or in a late fusion approach and found that both improved the performance of the emotion recognition model compared to unimodal approaches with a slightly higher accuracy for early fusion. Later, the authors presented a multimodal emotion recognition framework called EmotionMeter that also combines EEG and eye tracking data to recognize emotions in real-world applications. They successfully classified four different emotions with an accuracy of more than 85% using a multimodal neural network, outperforming both single modalities (Zheng et al., 2019). Another study on multimodal emotion recognition was conduced by López-Gil et al. (2016) who found that combining different signal sources on the feature level enables the detection of self-regulatory behavior more effectively than only using EEG data. Most recently, Wu et al. (2021) fused EEG and eye tracking data for emotion classification using effective deep learning for a gradient neural network. They report an 88% accuracy for the recognition of eight emotions.

The authors of Zhu et al. (2020) demonstrated that when eye movement and EEG data are combined for the detection of depression, a content-based ensemble method outperforms traditional approaches. The mental workload level is another cognitive state that can be classified using the proposed multimodal data. Debie et al. (2021) state in their review, that the combined features outperform single modalities for workload assessments. For example, Lobo et al. (2016) fused previously extracted eye tracking and EEG features on the feature level before training person-dependent and person-independent classifiers on them. They found that an almost perfect classification performance could be achieved for individual classifiers while independent classifiers only reached a lot worse accuracy.

1.3. Attentional State Classification

This study will examine different feature fusion strategies for a multimodal classification of EEG and eye tracking data to recognize internally and externally directed attention in a paradigm that manipulates internal/external attention demands. In general, attentional mechanisms are applied to filter the vast amount of available information at every moment for a better focus on relevant goals. Internally directed attention refers to a focus on information that is independent of sensory input, such as thoughts, memories, or mental arithmetic. It can occur deliberately (e.g., planning; Spreng et al., 2010) or spontaneously (e.g., mind wandering; Smallwood and Schooler, 2006). Externally directed attention instead describes a state of attentiveness to sensory input produced by the surroundings (Chun et al., 2011). Because concurrent self-evaluation of attentiveness to internal/external states while completing particular tasks would directly interfere with the direction of attention itself, a common approach is to ask participants in retrospect. Arguably, a system that would concurrently monitor the attentional state without interfering with the user may be better suited for application.

The suitability of eye tracking data for this classification task was shown by Annerer-Walcher et al. (2021) who achieved a classification accuracy of 69% for 4 s windows of raw eye tracking data. They compared gaze-specific properties and found that blinks, pupil diameter variance, and fixation disparity variance indicated differences in attentional direction. In Putze et al. (2016) and Vortmann et al. (2019a), the authors showed that such attentional differences can also be classified from EEG in different settings. They achieved 74.3% for 2 s windows and 85% for 13 s windows, respectively.

Eye tracking and EEG data have been collected simultaneously in several studies on attention (e.g., Vortmann and Putze, 2021). Kulke et al. (2016) investigated neural differences between covert and overt attention using EEG. The eye gaze was analyzed to control the correct labeling of the data. Dimigen et al. (2011) performed a co-registration of eye movement and EEG data for reading tasks and analyzed the fixation-related potentials. However, in these studies, the modalities were not combined but used for different purposes during the analysis.

To the best of our knowledge, the only paper that addresses feature fusion of EEG and eye tracking data for internally and externally directed attention in the context of attention classification is by Vortmann et al. (2019b). The authors implemented a real-time system for the attentional state classification and found that a late fusion approach with a decision rule improves the classification result of both single modalities. For 1.5 s data windows, the classification accuracy for the EEG data ranged between 0.56 and 0.81, for eye tracking data between 0.46 and 0.78 and for the late fusion approach between 0.58 and 0.86, calculated for 10 participant and a chance level of 0.5.

This work will systematically compare the unimodal approaches for EEG and eye tracking data with early, middle, and late fusion multimodal approaches for internally and externally directed attention.

2. METHODS

A dataset of 36 participants was analyzed for within-person classification accuracies of different multimodal neural networks.

2.1. Data

The data used in this study was recorded by Ceh et al. (2020)[1]. It encompasses EEG and eye tracking recordings of 36 participants (24 female, 12 male; age: M = 24 SD = 2.72; all right-handed; four had corrected-to-normal vision). The data set was chosen because the EEG and the eye tracking data were sampled with the same sampling rate. This makes the temporal alignment for the early fusion approaches easier and more accurate. The data collection was performed in a controlled laboratory setup which results in higher quality data and less confounding factors compared to more flexible setups that require, for instance, free movements (Vortmann and Putze, 2020).

[1] Publicly available at 10.17605/OSF.IO/5U6R9.

2.1.1. Task

During the recording, the participants had to perform two different tasks under two different conditions each. For all tasks, a meaningful German word of four letters was presented. For one task, the participants had to create **anagrams** of the word (i.e., "ROBE" is transformed to "BORE"). For the other task, a four-word long **sentence** had to be generated, each word starting with one of the four letters from the presented word (i.e., "ROBE" is transformed to "Robert observes eye behavior"). The employed paradigm builds on both a convergent (anagram) and divergent (sentence generation) thinking task and has been used in several studies investigating the effect of attention demands in the visual domain (Benedek et al., 2011, 2016, 2017; Ceh et al., 2020, 2021). Within the tasks, the attentional demands are manipulated using stimulus masking: in half of all trials, the stimulus is masked after a short processing period (500 ms), requiring participants to keep and manipulate the word in their minds. This enforces completion of the task relying on internally directed attention. In the other half of all trials, the stimulus word is continuously available (20 s), allowing for continuous retrieval using external sensory processing. The paradigm thus differentiates convergent and divergent thinking in a more internal vs. external attentional setting. For a detailed description of the task, see the original article.

2.1.2. Conditions

The effects of manipulating attention using these tasks were previously looked at for EEG (Benedek et al., 2011), fMRI (Benedek et al., 2014), and eye tracking (Benedek et al., 2017) data, or a combination of EEG and ET (Ceh et al., 2020), and fMRI and eye tracking (Ceh et al., 2021) data. Across these studies, the investigators found robust differences between the internal and external conditions on the level of eye behavior (e.g., increased pupil diameter during internally directed cognition; Benedek et al., 2017; Ceh et al., 2020, 2021), EEG (e.g., relatively higher alpha power over parieto-occipital regions during internally directed cognition; Benedek et al., 2011; Ceh et al., 2020), and fMRI (e.g., internally directed cognition was associated with activity in regions related to visual imagery, while externally directed cognition recruited regions implicated in visual perception; Benedek et al., 2016; Ceh et al., 2021). The observed attention effects were highly consistent across both tasks in all studies (i.e., across different modalities).

In this study, we will not differentiate between the two tasks. The classification will be based on masked (internally directed attention) and unmasked (externally directed attention) stimuli. Each participant performed 44 trials of each condition (chance level for the classification = 0.5).

2.1.3. Recordings

EEG was recorded with a BrainAmp amplifier by Brain Products GmbH with a sampling rate of 1,000 Hz using 19 active electrodes, positioned according to the 10-20 system in the following positions: Fp1, Fp2, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, O1, and O2. Additionally, three electrooculogram electrodes were included (left and right of the eyes, and adjacent to the radix nasi). References were placed on the left and right mastoid and the ground electrode was placed centrally on the forehead. Impedances were kept below 30 kOhm.

The eye tracking data was recorded using an EyeLink 1000 Plus eye tracker by SR Research Ltd. with a sampling rate of 1,000 Hz. For a more detailed description of the experimental setup and procedure (see Ceh et al., 2020).

2.2. Preprocessing

Simple preprocessing steps were applied to both data input sets to reduce the noise in the data. The classification will be performed per participant, with participant-dependently trained classifiers. Thus, correcting data to account for inter-individual differences is not necessary.

For the **eye tracking**, the X- and Y- coordinates and the pupil diameter of the left and the right eye were cleaned from non-existing values by dropping the respective samples. Binocular blinks (as defined by the eye tracker's built-in detection algorithm) were also excluded. The X- and Y-coordinates recorded by the eye tracker can be interpreted as the current gaze position relative to the screen.

The **EEG data** were processed using the MNE toolbox by Gramfort et al. (2013). First, the data was bandpass-filtered between 1 and 45 Hz using windowed FIR filters. An additional notch filter was applied at 50 Hz (power-line noise). Afterward, the data was re-referenced to average. Bad channels or epochs were not excluded from the data.

For both data sets, each trial was cut into four non-overlapping 3 s windows: 3–6, 7–10, 11–14, and 15–18 s after trial onset. The first seconds of each trial were not used to avoid an effect of the masking process in the data. In total, each participant's data set contained $4^*44 = 176$ data windows. No baseline correction was applied.

We generated two **feature sets** for each modality. As argued earlier, for early feature fusion approaches, the input format from both modalities must be temporally compatible so it can be combined. The data synchronization was performed on the basis of the available timestamps. Missing values were dropped for both modalities. The first feature set is the plain **preprocessed time series**, without any further computations or feature extraction steps. This raw input has been proven suitable for EEG data classification (Schirrmeister et al., 2017). To generate the second feature set, we followed an approach introduced in Wang and Oates (2015). The authors suggest transforming time-series data into **representative images** that convolutional neural networks can classify. The first algorithm for the image generation is called Markov Transition Field (MTF). MTFs represent transition probabilities between quantiles of the data. As a second algorithm, they suggest Gramian Angular Summation Fields (GASF), which visualizes the distances between polar-coordinates of the time series data. They argue that both approaches keep spatial and temporal information about the data. The application of this feature generation approach for eye tracking data during internally and externally directed attention was implemented by Vortmann et al. (2021). They were able to show that the imaging time-series approach with a convolutional neural net achieve higher classification accuracies than classical eye gaze-specific features.

TABLE 1 | Shallow FBCSP Convolutional Neural Network structure (shallow FBCSP CNN) from Schirrmeister et al. (2017), implemented using the braindecode toolbox by Schirrmeister et al. (2017).

Layer name	Type	Properties
conv_time	Conv2d	Out = 40, kernel_size = (25, 1), stride = (1, 1)
conv_spat	Conv2d	Out = 40, kernel_size = (1, 23), stride = (1, 1)
bnorm	BatchNorm2d	Out = 40, eps = 1e-05, momentum = 0.1
pool	AvgPool2d	Kernel_size = (75, 1), stride = (15, 1), padding = 0
drop	Dropout	$p = 0.5$
conv_classifier	Conv2d	Out = 2, kernel_size = (194, 1), stride = (1, 1)

TABLE 2 | Simple Convolutional Neural Network structure (simple CNN) similar to Vortmann et al. (2021), implemented using the PyTorch library by Paszke et al. (2019). fc, fully connected.

Layer name	Type	Properties
conv1	Conv2d	Out = 60, kernel_size = (5, 5), stride = (1, 1)
conv2	Conv2d	Out = 120, kernel_size = (5, 5), stride = (1, 1)
conv_dropout	Dropout2d	$p = 0.5$
fc1	Linear	In = 9,720, out = 500
fc2	Linear	In = 500, out = 120
fc3	Linear	In = 120, out = 20
fc4	Linear	In = 20, out = 2

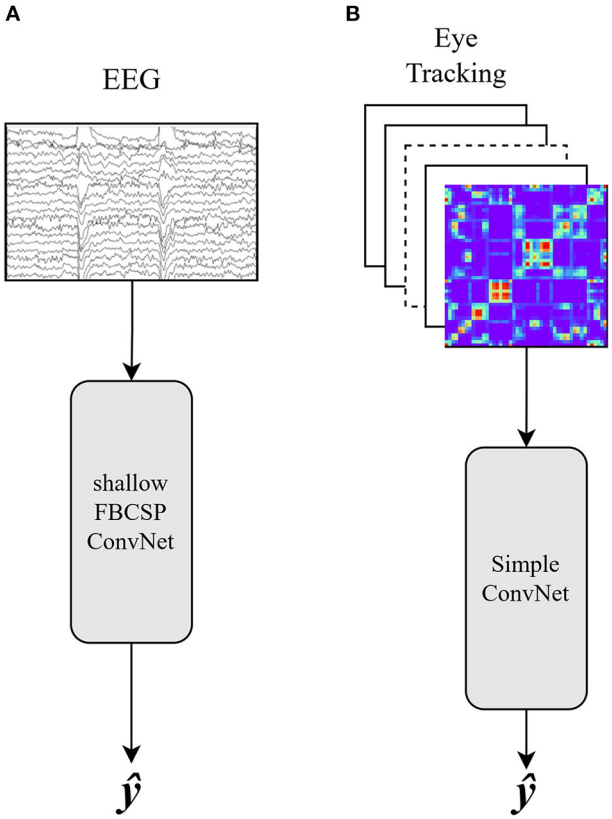

FIGURE 1 | (A) Single modality classification approach for EEG data using the shallow FBCSP Convolutional Neural Network. The preprocessed time-series is used as the input. **(B)** Single modality classification approach for eye tracking data using a Simple Convolutional Neural Network. The stacked images represent the Markov Transition Fields and Gramian Angular Fields computed on the raw data (see section 2.2). Both: $\hat{y}$ is the class prediction.

We calculated the MTF and the GASF image with 48x48 pixels for each channel in the data, resulting in 12 images for the eye tracking data: 2 images $*$ 2 eyes $*$ [x-coordinate, y-coordinate, pupil diameter] and 44 images for the EEG data: 2 images $*$ (22 EEG channels + 3 EOG channels). This results in an image matrix of 56 images per trial.

2.3. Classifier

The classification was performed in a person-dependent manner, resulting in an individual model for each participant. We used two different convolutional neural networks as classification algorithms, one for each feature set (time-series features and image features). Schirrmeister et al. (2017) introduced a shallow CNN that was inspired by Filterbank Common Spatial Pattern (FBCSP) analysis for EEG time-series. The layers of the network can be seen in **Table 1**. This **shallow FBCSP CNN** will be used to classify the time series feature set of both modalities. As optimizer, we used the AdamW optimizer (Loshchilov and Hutter, 2017), null loss, a learning rate of $0.0625 * 0.01$, and a weight decay of $0.5*0.0001$.

The second neural network that we used for the image features was the **simple CNN** adapted from Vortmann et al. (2021). **Table 2** describes the network structure in detail. This time, the Adam optimizer (Kingma and Ba, 2014), cross-entropy loss, a learning rate of 0.0001, and no weight decay were used. The label prediction the maximum of the softmax of the output layer was calculated.

In the first step, we classified the data using single modality approaches. The data were randomly split into training and testing data, using 33% for testing (stratified). We trained for a maximum of 30 epochs with a batch size of 40. Early stopping was applied if the classification accuracy on the training data was above 95% for more than five epochs to avoid overfitting.

The EEG data were classified using the time series feature set and the shallow FBCSP CNN (see **Figure 1A**). The eye tracking data were classified using the image feature set and a simple CNN (see **Figure 1B**). All evaluations are based on the network accuracy tested on the test data. Because of the equal distribution of the two conditions, the chance level for a correct window classification is 50%. The training and testing split, followed by the classification process, was repeated five times for each participant with each modality and fusion approach. As a final result for each participant, we calculated the average accuracy for the five runs.

2.4. Fusion Approaches

We compared the single modality results to four different fusion approaches. For the early feature fusion, we implemented two different versions: (1) the image feature sets of the EEG and eye tracking data are concatenated and classified by a simple

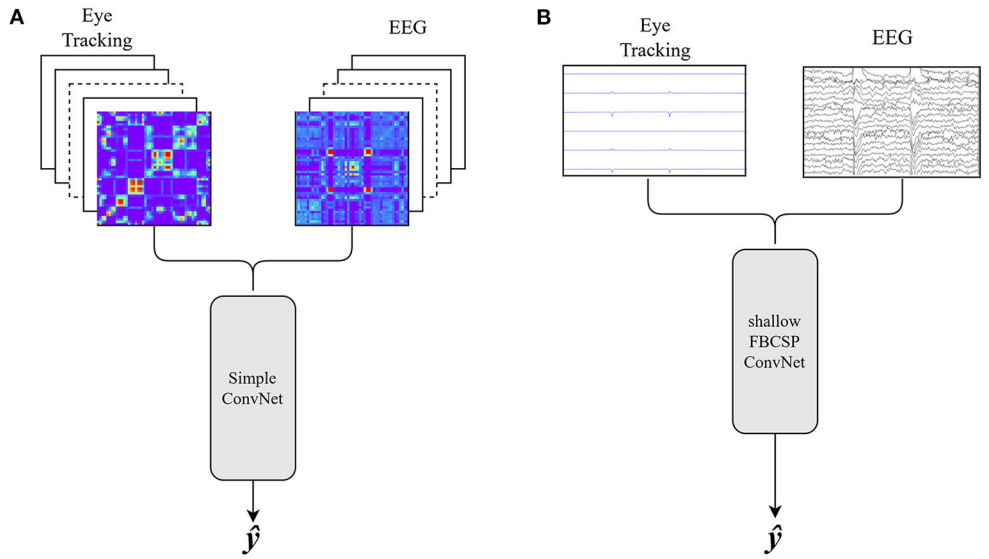

FIGURE 2 | (A) Early fusion approach using image features for EEG and eye tracking data. The stacked images are classified by a Simple Convolutional Neural Network. **(B)** Early fusion approach using time series of EEG and eye tracking data. The combined input is classified by the shallow FBCSP Convolutional Neural Network. Both: $\hat{y}$ is the class prediction.

CNN, and (2) the time series feature sets of both modalities are combined and classified using the shallow FBCSP CNN (see **Figure 2**). All parameters and training strategies were identical to the single modality classification process described in section 2.3.

In the middle fusion approach, the time-series features of the EEG data and the image features of the eye tracking data were used. As described in **Figure 3**, both feature sets were first processed simultaneously by different neural networks. A reduced version of the shallow FBCSP CNN got trained on the EEG data. The reduced model is identical to the model described in **Table 1** but the output size of the last layer (conv_classifier) was increased to 40. The eye tracking data were used to train the first layers of a simple CNN, until after the first linear layer (fc1; see **Table 2**). At this point, the outputs of both networks got concatenated, changing the input size of the second fully connected layer (fc2) before passing through the rest of the linear layers of a simple CNN.

Lastly, in the late fusion approach, the EEG and eye tracking data were classified separately as described for the single modality approaches. The prediction probabilities of both classes were used to decide on the final prediction (see **Figure 4**). We used the following decision rule: if both modalities predict the same label, use it as the final prediction. Else, if the probability of the EEG prediction $P(\hat{y}) > 0.5$, use the label predicted by the EEG classifier. Else, use the label that was predicted by the eye tracking classifier.

The decision was mutual (case 1) in 0.572 ± 0.074 of the trials. For 0.368 ± 0.071 of the trials, the EEG prediction was passed on and for 0.06 ± 0.024 the eye tracking decision was used.

3. RESULTS

All reported results are the statistics computed across all participants. We will first report the mean, standard deviation, range, and 95%-confidence interval of each approach, before testing for significant differences. All results can be seen in **Figure 5**.

The EEG-based single unimodal classification reached an average accuracy of 0.635 ± 0.095. The results ranged from 0.450 to 0.859, and the 95%-confidence interval of the classification accuracy for a new subject is [0.603, 0.668].

For the eye tracking approach, the average accuracy was 0.582 $\pm$ 0.092 within the range [0.397, 0.870]. The 95%-confidence interval was [0.551, 0.614].

When both modalities were represented by their time-series and processed with the shallow FBCSP CNN (Early Fusion—TS), the mean accuracy was 0.572 ± 0.077 (range [0.386, 0.853]). With a 95% confidence, the classification accuracies for this approach will reach between 0.545 and 0.598. The early fusion approach using image features (Early Fusion—Images) reached an average accuracy of 0.608 $\pm$ 0.083 over all participants. The range for this approach was [0.422, 0.887] and the 95%-confidence interval [0.580, 0.636].

For the middle fusion, the mean accuracy was 0.617 $\pm$ 0.101, range of [0.431, 0.870], and 95%-confidence interval of [0.583, 0.652].

Finally, the late fusion approach with the decision rule described in section 2.4 achieved the highest mean classification accuracy with 0.642 $\pm$ 0.096, a range of [0.456, 0.881] and a confidence interval between 0.609 and 0.675.

We performed the significance analysis using a paired two-tailed t-test of the accuracy on all combinations of approaches (see **Table 3**). Our main aim in this study was to identify promising approaches for the feature combination of a multimodal classifier. These results hint at which approach is worth improving, adjusting, and optimizing further. Thus, we would prefer a False Positive over a False Negative because it would make us "exclude" a promising approach for further

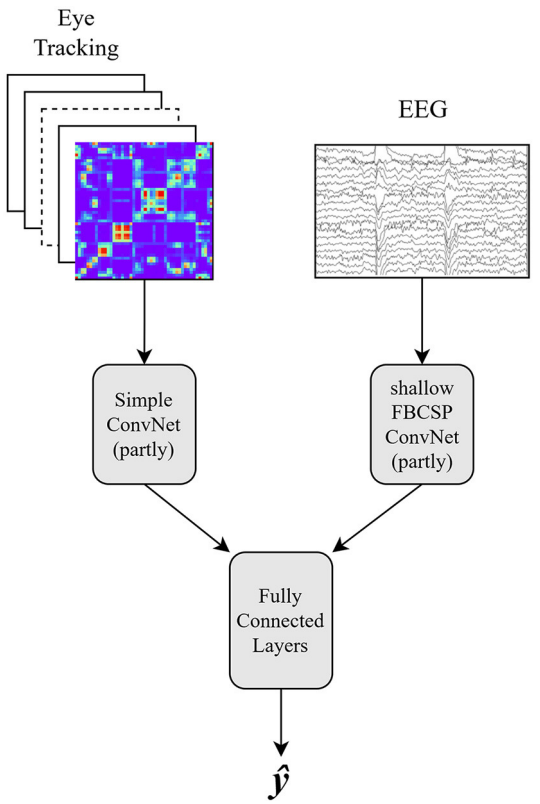

FIGURE 3 | Middle fusion approach using image features for eye tracking data in combination with the first steps of a Simple Convolutional Neural Network and the time series of EEG data with the reduced shallow FBCSP Convolutional Neural Network. The concatenated outputs are put through further linear layers. $\hat{y}$ is the class prediction.

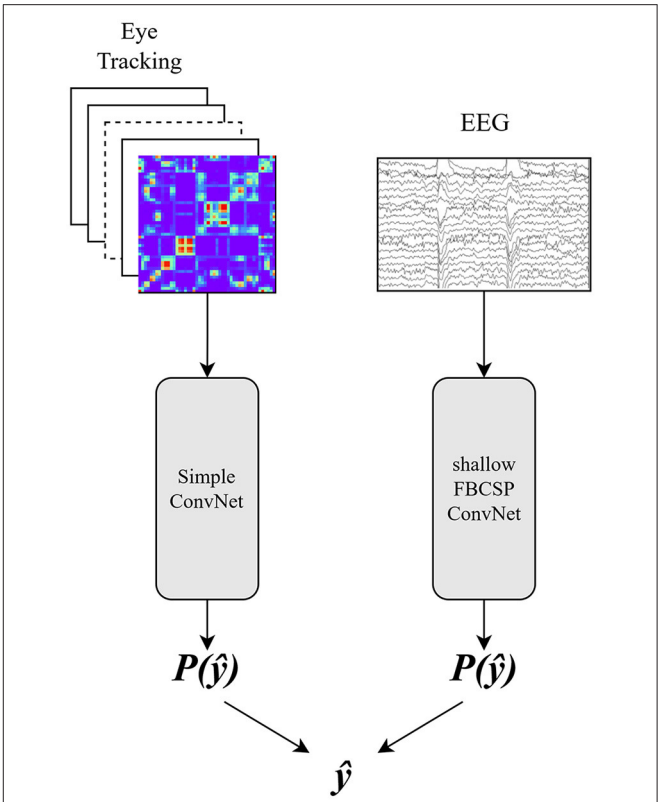

FIGURE 4 | Late fusion approach using image features for eye tracking data in combination with a Simple Convolutional Neural Network and the time series of EEG data with the shallow FBCSP Convolutional Neural Network. The probabilities of the predictions $P(\hat{y})$ are used for the final prediction $\hat{y}$.

studies on this topic. Following this philosophy, we chose a less conservative correction for multiple testing. By controlling the False Detection Rate (FDR) following Benjamini and Hochberg (1995), we find six significant differences. For the single modalities, the results for the EEG classification are not significantly better than the eye tracking results because they were identified as a false positive. Between the two early fusion approaches, the results obtained by the image feature set were significantly better than for the time-series features. No classification approach was significantly different from all other approaches, but the multimodal late fusion outperformed both unimodal classification approaches.

4. DISCUSSION

A system requires information in order to adapt more effectively to the needs of its users. The synergy may increase further, if a user does not have to explicitly state such requirements. Biosignals are a means of implicitly acquiring information, and combining multiple signals concurrently may result in a more accurate fit. Thus, we classified attention as internally or externally directed using 3 s multimodal EEG and eye tracking data in the current study. We compared different feature sets and feature fusion strategies. For the two feature sets and neural networks, we chose one combination that was previously used for EEG data (Schirrmeister et al., 2017) and one combination that was previously used for eye tracking data (Vortmann et al., 2021).

In a preliminary analysis of classification accuracies for the two single modalities, we discovered that prediction accuracies based on EEG data ($M = 0.635$) were significantly higher than those based on eye tracking data ($M = 0.582$). Regardless of the suitability of the modalities themselves, the disparities could also be explained by the disparate classification processes.

Interestingly, fusion of image features ($M = 0.608$) outperformed time series classification ($M = 0.572$) significantly for the two early fusion approaches. The image features were previously used for the eye tracking classification. As a result, we conclude that the different accuracies cannot be attributed solely to the quality of the classification approaches themselves. Rather than that, it appears as though the classification strategy and modality being used interact.

Neither of the early fusion approaches outperformed the single modalities by a significant margin. The time-series-based early fusion approach ($M = 0.572$) performed even worse than the unimodal EEG classification ($M = 0.635$). As discussed in the related work, other early fusion strategies have been used in the

TABLE 3 | *P*-values of two-tailed paired *t*-test for the comparison of the feature fusion approaches.

	EEG	ET	Early—TS	Early—images	Middle	Late
Only ET	**<0.001**					
Early fusion—TS	**= 0.005**	= 0.578				
Early fusion—images	= 0.141	= 0.068	**= 0.038**			
Middle fusion	= 0.42	= 0.091	**= 0.002**	= 0.648		
Late fusion	**=0.016**	**<0.001**	**= 0.003**	= 0.0693	= 0.268	
Average accuracy (%)	63.5	58.2	57.2	60.8	61.7	64.2

Significant differences are marked in bold. A significance threshold of $\alpha < 0.05$ is assumed. FDR correction by Benjamini and Hochberg (1995) was applied to correct for multiple testing. TS, time series.

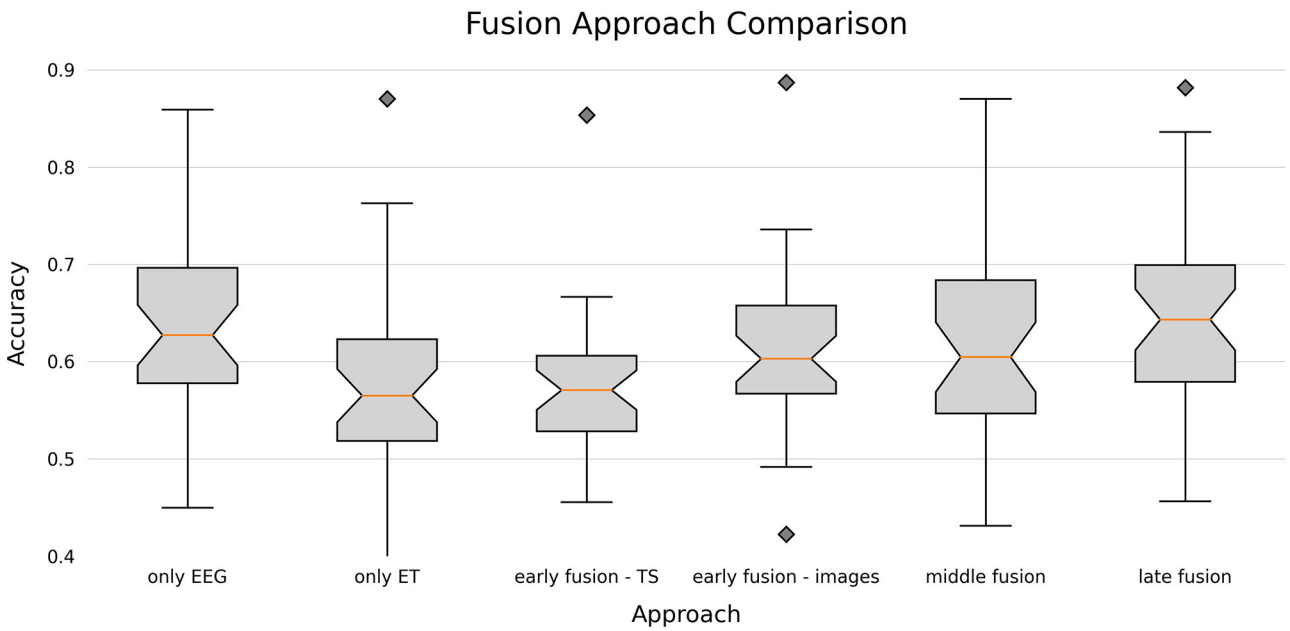

FIGURE 5 | Boxplots for the classification accuracies of the different fusion approaches visualizing the lower to upper quartile values of the data, with a line at the median. The whiskers represent the 95% range of the data. Diamonds represent outliers.

past to combine EEG and eye tracking data (Mangai et al., 2010; Liu et al., 2016; Guo et al., 2019). Different feature extraction algorithms or early statistics-based feature fusion techniques could be used in future studies to improve classification accuracy for the early fusion approaches. However, it was already noted in Polikar (2006) that early fusion is not reasonable as opposed to late fusion because of the diversity in the data. Thus, we see an advantage for middle and late fusion approaches.

As proposed in the section 1, a middle fusion could be an effective way to combine the advantages of feature-level and decision-level fusion. Individual modalities are processed independently first, resulting in classifier branches that are optimally adapted and trained for each modality. The two branches are connected in the middle, and the available data from both modalities can be used to train the rest of the network. While this approach enables correlations to be exploited, it also identifies significant unimodal data patterns that would be missed by other feature extraction approaches used in early fusion strategies. The primary difficulty with the middle fusion approach is network design. While it combines the strengths of the other two fusion strategies, it also incorporates their challenges. In a first step, suitable feature extraction and representation, as well as network structure for each modality, have to be found. These neural network branches must be designed in such a way that they allow for concatenation at a predetermined point. Finally, the neural network's subsequent layers must be appropriately designed for the merged modalities. On the one hand, complex correlations, and interactions must be discovered in order for the network to outperform a late fusion approach. On the other hand, the network's complexity must remain reasonable in comparison to the amount of data available. Otherwise, middle fusion networks will almost certainly have an excessively large number of parameters, rendering them unsuitable for a wide variety of applications.

It is difficult to generalize the results of the middle fusion, in particular: The neural network's structure is extremely adaptable, with an infinite number of possible configurations. The fully connected layers add parameters for successfully classifying multimodal data by learning correlations. The results of this study indicate that middle fusion is more promising than unimodal and early fusion approaches, but does not outperform late fusion. We assume that the network structure chosen was not optimal for maximizing the benefits of intermediate fusion. The layers were designed to resemble the individual unimodal networks and merged appropriately to maintain comparability. We hypothesize that more conservative and informed neural network engineering could significantly improve classification results. On the downside, this engineering is likely to be highly dataset and application dependent and will require a thorough understanding of the modalities' interactions.

In conclusion, our findings indicate that performing feature fusion in the middle of the classification process can slightly improve classification performance when compared to early fusion approaches. But supposedly, the neural network that intermediately combines the two modalities is subject to many adjustments and requires special engineering for each feature set combination and application.

While there was no significant difference between the middle fusion ($M = 0.617$) and the late fusion ($M = 0.642$), the late fusion approach was the only approach to significantly outperform both unimodal approaches in this data set. However, it did not outperform both early fusion approaches.

By comparison, the late fusion approach's optimization of the decision rule contains fewer parameters and is easily adaptable to new feature sets. However, the approach suggested here required expert knowledge to come up with a decision rule. For more efficient decision level fusion, statistical approaches or attention mechanisms could be applied (Mirian et al., 2011).

Improved unimodal classification pipelines would be a primary goal of improving late fusion. The primary disadvantage of the late fusion approach discussed in section 1 is the absence of correlation exploration between the modalities, which are processed independently. Thus, any information encoded in the early combination cannot be discovered using late fusion approaches that combine the modalities only at the decision level. A possible solution to this issue would be to add another "branch" of classification that predicts an output based on fused input, while maintaining the single modality classification. In our example, the decision rule would consider the EEG, ET, and a third combined prediction in addition to the two predicted labels and their probabilities.

We discovered during the training process that classification accuracy was highly dependent on the current training and test split for the same data set. Increasing the size of the data set may eliminate this effect. If more training data were available, the variance in the data would help to reduce bias and the likelihood of overfitting on the training data.

Another aspect that requires further thought is the inter-subject variability. The appropriate classification approach may depend on the participant and the quality of the data of each modality. For subjects with low individual EEG and eye tracking classification accuracies a middle or early fusion approach might increase the accuracy significantly. On the other hand, it the individual classification accuracies are already good, a late fusion might benefit from the modality specific classification.

We used a designated EEG and eye tracking co-registration study to have similar data quality for both modalities. The data was collected in a controlled laboratory environment. Applications and use cases with a more flexible setup and varying data qualities require another examination because one of the suggested approaches could be better suitable to correct for the worse quality of one modality than the others.

Overall, the differences between the approaches are not substantial enough to generally recommend the use of one over the others. We were able to show that a classification of more strongly internally vs. externally directed attention based on short data windows is possible above chance level for several approaches. We assume that the best fusion approach is highly dependent on the structure of the available multimodal data (e.g., sampling rate, data quality) and conclude that testing several approaches is necessary to find the most suitable for the data set. **Table 4** summarizes the advantages, challenges, and premises for each fusion approach.

TABLE 4 | Summary of the advantages, challenges, and premises for each fusion approach.

Fusion approach	Advantages (+), Challenges (–), and Premises (*)
Early fusion	+ Possibly finds correlations between modalities
	– Very different data structures to combine
	– Must use similar feature structures for all modalities
	* The same sampling rates for the data
	* Or preprocessing to adapt the data to each other
	* Best used when high chance of important modality interactions
Middle fusion	+ Tailored initial modality specific layers
	+ Possibly finds correlations between modalities
	+ Can work with different feature structures
	– Advanced NN engineering
	* Enough data for complex NN structure
	* Preliminary individual engineering of individual modalities was very different
	* Possibly important modality interactions
Late fusion	+ Tailored modality specific network design and features
	+ Missing data from one modality can be easily compensated
	– Finding a suitable decision rule or algorithm
	* Either good insight to find decision rule
	* Or enough data to train decision using ML
	* Best used when low chance of important modality interaction

4.1. Future Work

The current results may inspire further, more fine-grained comparisons even within the groups of early fusion approaches,

middle fusion networks, and late fusion decisions. On top of the presented suggestions on improving the current approaches, classification accuracy might increase if pre-trained models or transfer learning were applied For future work, other comparable data sets will be used to enlarge the data available for the training. The generalizability of the presented results should also be tested with further unrelated data sets. This study exclusively analyzed the data person-dependently. In the future, person-independence should be evaluated. The classification of unseen participants would include training the model on a pooled dataset of other participants, for example, in a leave-one-out approach. While the increased size of the training dataset might improve the accuracy of the classifier, the differences between participants might increase the variance in the dataset. Previous results have shown that the person-independent classification of EEG data is difficult and person-specific models are still the norm (Vortmann and Putze, 2021), whereas attempts to classify the eye tracking data of unseen participants for different attentional states were promising on larger datasets (Vortmann et al., 2021). However, the problem of generalizability was already discussed by Annerer-Walcher et al. (2021) who state that for internally and externally directed attention eye tracking data does not generalize well over participants. Our results have shown that a multimodal classifier outperforms unimodal classifiers for within-person training and testing and the next step will be to explore whether these improvements also hold for person-independent classification. For the real-time application of such a classifier in a BCI, the possibility to classify unseen participants without the need for person-dependent training data would highly increase the range of applications and the usability.

DATA AVAILABILITY STATEMENT

Materials and data are provided on the Open Science Framework (OSF, https://osf.io/5u6r9/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Graz, Austria. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

L-MV performed the analysis and prepared the manuscript. SC and FP contributed to the final version of the manuscript. FP supervised the project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Austrian Science Fund (FWF): P29801.

REFERENCES

Annerer-Walcher, S., Ceh, S. M., Putze, F., Kampen, M., Körner, C., and Benedek, M. (2021). How reliably do eye parameters indicate internal versus external attentional focus? *Cogn. Sci*. 45, e12977. doi: 10.1111/cogs.12977

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Benedek, M., Bergner, S., Könen, T., Fink, A., and Neubauer, A. C. (2011). EEG alpha synchronization is related to top-down processing in convergent and divergent thinking. *Neuropsychologia* 49, 3505–3511. doi: 10.1016/j.neuropsychologia.2011.09.004

Benedek, M., Jauk, E., Beaty, R. E., Fink, A., Koschutnig, K., and Neubauer, A. C. (2016). Brain mechanisms associated with internally directed attention and self-generated thought. *Sci. Rep*. 6, 1–8. doi: 10.1038/srep22959

Benedek, M., Schickel, R. J., Jauk, E., Fink, A., and Neubauer, A. C. (2014). Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia* 56, 393–400. doi: 10.1016/j.neuropsychologia.2014.02.010

Benedek, M., Stoiser, R., Walcher, S., and Körner, C. (2017). Eye behavior associated with internally versus externally directed cognition. *Front. Psychol*. 8:1092. doi: 10.3389/fpsyg.2017.01092

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Ceh, S. M., Annerer-Walcher, S., Körner, C., Rominger, C., Kober, S. E., Fink, A., et al. (2020). Neurophysiological indicators of internal attention: an electroencephalography-eye-tracking coregistration study. *Brain Behav*. 10, e01790. doi: 10.1002/brb3.1790

Ceh, S. M., Annerer-Walcher, S., Koschutnig, K., Körner, C., Fink, A., and Benedek, M. (2021). Neurophysiological indicators of internal attention: an fMRI-eye-tracking coregistration study. *Cortex* 143, 29–46. doi: 10.1016/j.cortex.2021.07.005

Cheng, S., Wang, J., Zhang, L., and Wei, Q. (2020). Motion imagery-BCI based on EEG and eye movement data fusion. *IEEE Trans. Neural Syst. Rehabil. Eng*. 28, 2783–2793. doi: 10.1109/TNSRE.2020.3048422

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol*. 62, 73–101. doi: 10.1146/annurev.psych.093008.100427

Debie, E., Fernandez Rojas, R., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., et al. (2021). Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE Trans. Cybern*. 51, 1542–1555. doi: 10.1109/TCYB.2019.2939399

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and eeg in natural reading: analyses and review. *J. Exp. Psychol*. 140, 552. doi: 10.1037/a0023885

D'Mello, S., and Kory, J. (2012). "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12* (New York, NY: Association for Computing Machinery), 31–38. doi: 10.1145/2388676.2388686

Dong, X., Wang, H., Chen, Z., and Shi, B. E. (2015). "Hybrid brain computer interface via Bayesian integration of EEG and eye gaze," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)* (Montpellier), 150–153. doi: 10.1109/NER.2015.7146582

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci*. 7, 267. doi: 10.3389/fnins.2013.00267

Guo, J.-J., Zhou, R., Zhao, L.-M., and Lu, B.-L. (2019). "Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Berlin), 3071–3074. doi: 10.1109/EMBC.2019.8856563

Kim, M., Kim, B. H., and Jo, S. (2015). Quantitative evaluation of a low-cost noninvasive hybrid interface based on EEG and eye movement. *IEEE Trans. Neural Syst. Rehabil. Eng.* 23, 159–168. doi: 10.1109/TNSRE.2014.2365834

Kim, M., Lee, S., and Kim, J. (2020). "Combining multiple implicit-explicit interactions for regression analysis," in *2020 IEEE International Conference on Big Data (Big Data)* (Atlanta, GA), 74–83. doi: 10.1109/BigData50022.2020.9378402

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kulke, L. V., Atkinson, J., and Braddick, O. (2016). Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking. *Front. Hum. Neurosci.* 10, 592. doi: 10.3389/fnhum.2016.00592

Liu, W., Zheng, W.-L., and Lu, B.-L. (2016). "Emotion recognition using multimodal deep learning," in *International Conference on Neural Information Processing* (Kyoto: Springer), 521–529. doi: 10.1007/978-3-319-46672-9_58

Lobo, J. L., Ser, J. D., De Simone, F., Presta, R., Collina, S., and Moravek, Z. (2016). "Cognitive workload classification using eye-tracking and EEG data," in *Proceedings of the International Conference on Human-Computer Interaction in Aerospace, HCI-Aero '16* (New York, NY: Association for Computing Machinery). doi: 10.1145/2950112.2964585

López-Gil, J.-M., Virgili-Gomá, J., Gil, R., Guilera, T., Batalla, I., Soler-González, J., et al. (2016). Method for improving EEG based emotion recognition by combining it with synchronized biometric and eye tracking technologies in a non-invasive and low cost way. *Front. Comput. Neurosci.* 10, 85. doi: 10.3389/fncom.2016.00119

Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, Y., Zheng, W.-L., Li, B., and Lu, B.-L. (2015). "Combining eye movements and EEG to enhance emotion recognition," in *IJCAI, Vol. 15* (Buenos Aires), 1170–1176.

Mangai, U. G., Samanta, S., Das, S., and Chowdhury, P. R. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Techn. Rev.* 27, 293–307. doi: 10.4103/0256-4602.64604

Mirian, M. S., Ahmadabadi, M. N., Araabi, B. N., and Siegwart, R. R. (2011). Learning active fusion of multiple experts' decisions: an attention-based approach. *Neural Comput.* 23, 558–591. doi: 10.1162/NECO_a_00079

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems, Vol. 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver: Curran Associates, Inc.), 8024–8035.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Polikar, R. (2012). *Ensemble Learning*. Boston, MA: Springer US, 1–34. doi: 10.1007/978-1-4419-9326-7_1

Putze, F., Scherer, M., and Schultz, T. (2016). "Starring into the void? Classifying internal vs. external attention from EEG," in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg), 1–4. doi: 10.1145/2971485.2971555

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Smallwood, J., and Schooler, J. W. (2006). The restless mind. *Psychol. Bull.* 132, 946. doi: 10.1037/0033-2909.132.6.946

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W., and Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage* 53, 303–317. doi: 10.1016/j.neuroimage.2010.06.016

Vortmann, L.-M., Knychalla, J., Walcher, S., Benedek, M., and Putze, F. (2021). Imaging time series of eye tracking data to classify attentional states. *Front. Neurosci.* 15, 625. doi: 10.3389/fnins.2021.664490

Vortmann, L.-M., Kroll, F., and Putze, F. (2019a). EEG-based classification of internally-and externally-directed attention in an augmented reality paradigm. *Front. Hum. Neurosci.* 13, 348. doi: 10.3389/fnhum.2019.00348

Vortmann, L.-M., and Putze, F. (2020). "Attention-aware brain computer interface to avoid distractions in augmented reality," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu), 1–8. doi: 10.1145/3334480.3382889

Vortmann, L.-M., and Putze, F. (2021). Exploration of person-independent bcis for internal and external attention-detection in augmented reality. *Proc. ACM Interact. Mobile Wear. Ubiquit. Technol.* 5, 1–27. doi: 10.1145/3463507

Vortmann, L.-M., Schult, M., Benedek, M., Walcher, S., and Putze, F. (2019b). "Real-time multimodal classification of internal and external attention," in *Adjunct of the 2019 International Conference on Multimodal Interaction* (Suzhou), 1–7. doi: 10.1145/3351529.3360658

Wang, Z., and Oates, T. (2015). "Imaging time-series to improve classification and imputation," in *Twenty-Fourth International Joint Conference on Artificial Intelligence* (Buenos Aires: AAAI Press), 3939–3945. doi: 10.5555/2832747.2832798

Wu, Q., Dey, N., Shi, F., Crespo, R. G., and Sherratt, R. S. (2021). Emotion classification on eye-tracking and electroencephalograph fused signals employing deep gradient neural networks. *Appl. Soft Comput.* 110, 107752. doi: 10.1016/j.asoc.2021.107752

Zhang, W., Ji, X., Yang, Y., Chen, J., Gao, Z., and Qiu, X. (2018). "Data fusion method based on improved ds evidence theory," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (Shanghai), 760–766. doi: 10.1109/BigComp.2018.00145

Zheng, W.-L., Dong, B.-N., and Lu, B.-L. (2014). "Multimodal emotion recognition using EEG and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Chicago, IL), 5040–5043.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. (2019). Emotionmeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zhu, J., Wang, Z., Gong, T., Zeng, S., Li, X., Hu, B., et al. (2020). An improved classification model for depression detection using EEG and eye tracking data. *IEEE Trans. NanoBiosci.* 19, 527–537. doi: 10.1109/TNB.2020.2990690

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.