Int J Artif Intell Educ (2025) 35:823–857
https://doi.org/10.1007/s40593-024-00412-2

ARTICLE

From the Lab to the Wild: Examining Generalizability of Video-based Mind Wandering Detection

Babette Bühler[1] · Efe Bozkir[2,6] · Patricia Goldberg[1] · Ömer Sümer[3] · Sidney D'Mello[4] · Peter Gerjets[5] · Ulrich Trautwein[1] · Enkelejda Kasneci[6]

Accepted: 14 May 2024 / Published online: 17 June 2024

Abstract

Student's shift of attention away from a current learning task to task-unrelated thought, also called mind wandering, occurs about 30% of the time spent on education-related activities. Its frequent occurrence has a negative effect on learning outcomes across learning tasks. Automated detection of mind wandering might offer an opportunity to assess the attentional state continuously and non-intrusively over time and hence enable large-scale research on learning materials and responding to inattention with targeted interventions. To achieve this, an accessible detection approach that performs well for various systems and settings is required. In this work, we explore a new, generalizable approach to video-based mind wandering detection that can be transferred to naturalistic settings across learning tasks. Therefore, we leverage two datasets, consisting of facial videos during reading in the lab (N = 135) and lecture viewing in-the-wild (N = 15). When predicting mind wandering, deep neural networks (DNN) and long short-term memory networks (LSTMs) achieve F_1 scores of 0.44 (AUC-PR = 0.40) and 0.459 (AUC-PR = 0.39), above chance level, with latent features based on transfer-learning on the lab data. When exploring generalizability by training on the lab dataset and predicting on the in-the-wild dataset, BiLSTMs on latent features perform comparably to the state-of-the-art with an F_1 score of 0.352 (AUC-PR = 0.26). Moreover, we investigate the fairness of predictive models across gender and show based on post-hoc explainability methods that employed latent features mainly encode information on eye and mouth areas. We discuss the benefits of generalizability and possible applications.

Keywords Mind wandering · Transfer-learning · Affective computing · Educational technology

Introduction

Attention plays a central role in learning and knowledge construction (Levine, 1990). However, a recent meta-analysis by Wong et al. (2022) showed that about 30% of

Extended author information available on the last page of the article the time learners spend in educational activities, their thoughts are elsewhere. This shift of attention away from the current task to task-unrelated thought is called mind wandering (Smallwood & Schooler, 2006). Wong et al. (2022) further demonstrated that frequent occurrence of task-unrelated thoughts during learning is significantly associated with lower test performance and explains about 7% of the variability in learning outcomes. This negative relationship holds equally for surface and inference-level learning and is consistent across tasks. For instance, mind wandering has shown to have a negative effect on reading comprehension (Smallwood, 2011; Feng et al., 2013; D'Mello & Mills, 2021; Bonifacci et al., 2022; Caruso & D'Mello, 2023) and lecture retention (Risko et al., 2012; Szpunar et al., 2013; Hollis & Was, 2016; Pan et al., 2020).

This evident effect of mind wandering on learning should not be neglected. Therefore, learning environments - physical as well as online - aim to create appealing conditions that allow students to focus their attention on the relevant content and support successful learning. The Covid19 pandemic has greatly accelerated the use (Lemay et al., 2021) and development (Dhawan, 2020) of online learning tools at all levels of education. These include intelligent tutoring systems (ITSs), massive open online courses (MOOCs), as well as online lecture portals. To support learners in online learning settings, one can either improve the presented learning materials in a way that decreases mind wandering, for instance, by making texts more interesting (Bonifacci et al., 2022), or one can try to direct attention back to the learning task, for instance, through targeted interventions (e.g., Mills et al., 2021).

One step that is foundational to those two approaches to support learners is the automated detection of mind wandering, as it allows for continuous and unobtrusive measurement of the state of attention over time. It can be used to test and optimize learning materials and to conduct further research on the conditions under which mind wandering occurs and its effects on learning outcomes. At the same time, it offers online learning systems the possibility to implement targeted interventions to respond to the learners' shift of attention with adaptive and supporting actions. It has been demonstrated that such automated interventions can reduce mind wandering and thus support learning. For example, feedback following eye-tracking-based mind wandering detection mitigated its negative effect on reading comprehension during computerized reading (D'Mello et al., 2017; Mills et al., 2021). Furthermore, repetition and questioning interventions based on automatically detected mind wandering reduced mind wandering and improved retention of students with low prior knowledge in an ITS in certain cases (Hutt et al., 2021).

Mind wandering detectors using supervised machine learning mostly rely on data from modalities such as eye trackers (Hutt et al., 2017; Faber & D'Mello, 2018; Hutt et al., 2019; Faber et al., 2020; Zhang et al., 2020; Mills et al., 2021) or physiological sensors such as EEG (Jin et al., 2019; Dong et al., 2021). While these modalities provide very useful process information, the use of such sensors requires a well-controlled environment, is quite costly, and is difficult to scale. However, another strand of recent research has shown that mind wandering can also be detected above chance level using video recordings of the face obtained from consumer-grade webcams, such as those found in almost all laptops (Bosch & D'Mello, 2021; Lee et al., 2022). The use of video recordings enables the detection of mind wandering on a large scale in natural environments where online learning systems are commonly used, such as classrooms or homes.

Video-based mind wandering detectors have the potential to be used in many different systems and environments. Intelligent user interfaces for learning can combine multiple stimuli, such as text and video, and may be used globally, i.e., by culturally diverse target groups. However, to train suited machine-learning models, labeled ground truth data is needed, i.e., collecting learner self-reports makes the whole process time-consuming, effortful, and costly. Thus, approaches that generalize well across different settings, learning tasks, and target groups are required in order to ensure the applicability of such solutions.

To obtain generalizability, facial features rather than gaze features, which are highly predictive but also highly stimulus dependent (Faber et al., 2020), are suitable as they can achieve greater transferability between tasks. In this study, we use transfer-learning-based features trained on a dataset of facial expressions in the wild (Mollahosseini et al., 2017), i.e., on a highly diverse set of facial images. Previous research has shown that affective features such as facial action units (AUs) and predictions of emotional state are informative for predicting mind wandering (Stewart et al., 2017a; Bosch & D’Mello, 2021). However, this could pose a challenge when thinking about generalizing models across subject groups with different cultural backgrounds, as there is an ongoing scientific debate about the universality of facial expressions of emotion across different cultures (Russell, 1994; Ekman, 1994; Jack et al., 2012). This could have implications for cross-cultural generalizability when using features derived from such classification tasks. This highlights one of the major limitations of previous approaches-namely, the lack of sample diversity and the unexplored effects on algorithms, which may bias results and impact generalizability (Kuvar et al., 2023).

The goal of the present work is to examine the generalizability of video-based mind wandering detection. Towards this goal, we investigate whether (1) a new feature set based on transfer learning of facial-expression recognition can be used in combination with temporal models exploiting temporal relationships in video data to improve model performance compared to explicit facial features. Furthermore, (2) we explore the potential of its generalization across two datasets that differ with regard to the environment (lab vs. in the wild), task (reading vs. watching a video lecture), and cultural background of the target groups (American vs. Korean students). We then (3) examine the fairness of our models across genders and (4) use explainable AI tools to investigate the information encoded in latent features. Accurate detection of mind wandering is the first critical step towards large-scale research and adaptive learning technologies that aim to enhance engagement and learning outcomes.

Related Work

The automated detection of mind wandering episodes allows measuring this state non-obtrusively and continuously over time. This is achieved by employing machine learning methods and self-reports from learners as ground truth. Here, self-reports are used because an objective, reliable measurement by neurophysiological or behavioral markers is not possible so far (Smallwood & Schooler, 2015). When collecting self-reports with the probe-caught method, subjects are repeatedly interrupted by a probe and explicitly asked about the direction of their attention (Smallwood & Schooler, 2006), while in the self-caught method, participants are instructed to report whenever they become aware of their own shift of attention (Schooler et al., 2011). Research suggests that both approaches allow for reliable measurement of mind wandering (Schubert et al., 2020; Varao-Sousa & Kingstone, 2019). Self-reports are associated with physiological signals (Franklin et al., 2013; Blanchard et al., 2014; Christoff et al., 2009) and consistently correlate with objective performance measures (Randall et al., 2014), demonstrating predictive validity. Moreover, the datasets on which mind wandering is examined are all very unbalanced with approximately 25-30% mind wandering rates. Therefore, the results of the prediction are reported using the F_1 measure, which represents a harmonized mean of precision (proportion of predicted mind wandering instances that are truly mind wandering) and recall (proportion of true mind wandering instances predicted as mind wandering). The reported improvement over chance level is the proportion of above chance performance of the perfect above chance prediction. For more details on the evaluation metrics, we refer the reader to "Evaluation Metrics".

Most studies on automated mind wandering can be divided into two main strands according to the sensing modalities used for the detection: Eye tracking and physiological measures, and video recordings. Therefore, in the following, we review this research focusing on these areas and describe the novelty of our study.

Eye-tracking and Physiological Sensor-based Approaches

Most research on mind wandering detection has focused on eye movement data obtained by eye trackers. According to the so-called mind-eye link, cognitive processes are reflected in eye movements (Just & Carpenter, 1976; Rayner, 1998; Reichle et al., 2012) thus making eye tracking suitable to identify mind wandering. Global gaze features, such as fixations and saccades, as well as locality features describing the spatial properties of gaze have widely been used in research to estimate attentional states during a variety of learning-related tasks such as reading (Bixler & D'Mello, 2014; D'Mello et al., 2016; Faber et al., 2018; Mills et al., 2021), watching video lectures (Hutt et al., 2017; Jang et al., 2020; Zhang et al., 2020) or using an ITS (Hutt et al., 2019). Also cross-task prediction of such features was examined (Faber et al., 2020; Bixler & D'Mello, 2021). Additionally, pupil size and blink rates have been shown to be meaningful features to off-task thought detection (Smilek et al., 2010; Brishtel et al., 2020).

Further, several physiological sensors are utilized for mind wandering detection. Electrodermal activity (EDA) was used as a standalone modality (Blanchard et al., 2014) and in combination with eye tracking (Brishtel et al., 2020) for mind wandering detection during reading. Furthermore, an increased heart rate was detected in mind wandering episodes due to greater arousal (Smallwood et al., 2007), thus being deployed for automated detection as well (Pham & Wang, 2015). Another way of assessing mind wandering is using EEG (Jin et al., 2019; Dong et al., 2021; Dhindsa et al., 2019; Conrad & Newman, 2021), which has also been employed for learning related mind wandering during the watching of online lectures (Dhindsa et al., 2019; Conrad & Newman, 2021).

The collection of data employing eye trackers or physiological sensors requires highly controlled settings (i.e., laboratory). Consequently, most of the studies were conducted in a lab setting, with the exception of Hutt et al. (2019); who used commercial eye trackers in a classroom setting. Furthermore, respective modalities are often very expensive, which again limits scalability. The alternative of video-based detection is more cost-effective. Consumer-grade webcams can be employed and thus detection can be upscaled easily in naturalistic settings (e.g., at home or in the classroom). Therefore, in this study, we focus on mind wandering detection based only on facial videos.

Video-based Approaches

The first approach to detect mind wandering based on facial videos from webcams was provided back in 2017 by Stewart et al. (2017a). The authors predicted self-reported mind wandering during narrative film watching in a laboratory setting, based on features such as AUs, head pose, face position, face size, and gross body movement. Employing support vector machine (SVM) models, they achieved an F_1 score of 0.39, which is an improvement of 13% above chance level, on aggregated features from 45-second windows. In a following study, the potential of cross-task classification in laboratory settings was shown, by predicting mind wandering on a reading task from a model trained on mind wandering from a film watching task and vice versa (Stewart et al., 2017b). Employing the same feature set as in the previous study and a decision tree based C4.5 classifier, their models generalized well and almost maintained within-dataset prediction performance when training on film watching data and predicting reading data (F_1: 0.407; 21% above chance level) and also after adjusting the classification threshold the other way around (F_1: 0.441; 22% above chance level).

In a laboratory experiment to detect mind wandering during MOOCs, Zhao et al. (2017) implemented webcam-based gaze estimation and compared it to predictions with specialized eye-tracking data. With probe-caught mind wandering reports as ground truth they concluded that SVM classifiers on both data sources perform equally well with the webcam-based approach achieving an F_1 score of 0.405, a 16% above-chance improvement.

Another recent study by Bosch & D'Mello (2021) examined face-based mind wandering detection in the laboratory during a reading task with self-caught mind wandering reports and in the classroom during the usage of an ITS based on probe-caught mind wandering reports. In addition to AUs, head pose, and body movement, they hand-crafted new features depicting co-occurring AUs, temporal dynamics of AUs, and facial texture. In the classroom setting, those features were extracted in real-time, avoiding the recording of children due to privacy concerns. With SVM and deep neural networks on aggregated features sets over 10-second windows, they achieved F_1 scores of 0.478 in the lab and 0.414 in the classroom setting, which represents 25% and 20% above-chance improvements respectively. Although eye-tracking features (Global gaze features F_1: 0.45, 29% above-chance accuracy; Locality gaze features F_1: 0.49, 34% above-chance accuracy) outperform a facial-feature based approach

(AUs F_1: 0.31, 10% above-chance accuracy; Co-occurring AUs F_1: 0.3, 9% above-chance accuracy) for mind wandering detection while using an intelligent tutor system in the same classroom setting (Hutt et al., 2019), a fusion of both features could increase robustness by accounting for missing values in one of the two modalities.

In a recent paper by Lee et al. (2022) mind wandering detection based on facial webcam videos during lecture viewing in the wild, for example at home, was examined. Gaze-related features (i.e., speed, dispersion, horizontal movement ratio), Eye Aspect Ratio (EAR) features, head movement, as well as emotion predictions, were extracted from facial videos. They employed eXtreme Gradient Boosting (XGBoost), Deep Neural Network (DNN), and SVM classifiers on different time windows, achieving the best results with XGBoost with an F_1 score of 0.36 (15% over chance level improvement) utilizing 10-second windows.

Adopting webcam-based eye tracking for reading tasks, Hutt et al. (2023) executed two in-the-wild studies: the first recruited participants through a university, and the second utilized Prolific for participant recruitment. They predicted probe-caught mind wandering using global and local gaze features, achieving an F_1-score of 0.25 on a combined dataset (9% above chance). In cross-dataset prediction between two data collections, training on one dataset and predicting on the other, they achieved Kappa values of 0.09 and 0.15, respectively.

The overwhelming majority of these studies, except one cross-task laboratory-based generalization study (Stewart et al., 2017b) and a very recent webcam-eye-tracking study (Hutt et al., 2023), consider the performance of their models only on the labeled data on which they were trained and therefore do not test the generalizability of their models. As mentioned above, collecting labeled data for this specific task is costly and may not be possible for every use case, implying that generalizable models are necessary.

Novelty of this Work

While a large variety of machine learning and computer vision methods are used for feature extraction, they are not directly targeted to generalizability. Hence, there are still unexploited potentials of machine learning techniques, which we investigate in this study. These techniques include utilizing features from pre-trained networks for similar but more general tasks, employing temporal models, and an evaluation of generalizability, which are discussed as follows.

Deep Learning-based Features

To our knowledge, all previous research in this field used explicit features, such as AUs (Stewart et al., 2017a; Bosch & D'Mello, 2021) and gaze features (Lee et al., 2022) for predictions. In some cases, additional hand-manufactured features, such as gaze dispersion (Lee et al., 2022), have been created. While those appear to be informative in the data at hand, there is a risk that some of those features, especially gaze-related features, might be very stimulus-specific. Gaze in a reading task has a highly specific signature, whose features may be difficult to transfer to different tasks

(Bixler & D’Mello, 2021; Faber et al., 2020). While there have been advances with webcam-based eye tracking in recent years, enabling the detection of fixations of specific areas of interest (AOIs) on the computer screen, there are still severe limitations compared to specialized eye trackers. Especially in terms of mind wandering detection, more fine granular eye movement indicators such as saccade duration are predictive and most likely more generalizable. Those cannot be observed based on consumer-grade webcams at this time, underlining the advantages of facial expression features. Moreover, recent studies have showcased the utility of advanced Facial Emotion Recognition (FER) methods on webcam videos in adjacent domains, particularly for evaluating emotion regulation in remote collaborative learning settings (Nguyen et al., 2022a; Ngo et al., 2024). Further, the creation of hand-manufactured features requires domain knowledge and might be tailored to the setting at hand.

In other image-based classification tasks such as facial expression recognition, deep learning methods have been used successfully (Rawat & Wang, 2017; Li & Deng, 2020). Due to the limited sample sizes of data for mind wandering detection, we use transfer learning, where the feature extraction part of networks pre-trained on a similar task with available large datasets, is applied to a new problem. Based on the previous successful use of facial expression features as facial texture patches and AUs (Bosch & D’Mello, 2021) and emotional state features (Lee et al., 2022) for mind wandering detection, we argue that latent features learned by a CNN trained on the related task of FER can be informative to the mind wandering classification problem at hand. Furthermore, as they are trained on an in-the-wild data set, which assures large data variability, and thus utilizing these features may contribute to making the models generalizable to in-the-wild settings, i.e. in terms of image quality. To increase the confidence in these latent features, which cannot be directly interpreted by humans, we use the Explainable AI tool LIME (Ribeiro et al., 2016) to illustrate which parts of the face are used for classification and are encoded respectively in our latent representation.

Temporal Models

Previous research indicates that temporal dynamics can be informative for mind wandering detection. Recent studies showed the importance of hand-crafted features such as co-occurring AU pairs and temporal dynamics of AUs (Bosch & D’Mello, 2021), as well as body (Stewart et al., 2017a; Bosch & D’Mello, 2021) and head movement (Lee et al., 2022), which are explicitly designed to depict these dynamics. If not represented by such manually generated features, they might be blurred by aggregation over time. For this reason, we employ models that are able to take time-series data such as the video data at hand as input and directly learn temporal relations between features. The use of these temporal models allows us to additionally train an end-to-end model, in which we can fine-tune the pre-trained Convolutional Neural Network (CNN) used for frame-wise feature extraction to our specific classification task.

Generalizability

In previous face-based mind wandering detection research, only cross-task prediction in laboratory settings was explored (Stewart et al., 2017b), but the generalizability of lab settings to other more naturalistic settings is still to be examined. We trained a model on data collected during a reading task in the lab (Bosch & D'Mello, 2021) and then applied it to in-the-wild data of students watching a lecture video at home (Lee et al., 2022), resulting in a cross-context and cross-task prediction. However, the two data sets differ additionally in the cultural backgrounds of their subjects, as one was collected in the U.S. and the other in Korea. With regard to the discussion about the cross-cultural universality of facial expressions in the literature (Russell, 1994; Ekman, 1994; Benitez-Garcia et al., 2017), this is a further challenge for our model, which is based on such features to generalize across culturally diverse user groups. Although the in-the-wild dataset is relatively modest in size, it represents to our knowledge the only publicly available dataset of its kind, enabling a crucial initial stride towards achieving generalizability in naturalistic environments. For comparability, we used the reading-task data to carry out within-context and within-task evaluation. We further compare the classification results for both across gender to investigate potential biases.

Methodology

In this section, we discuss the details of the employed data, features that are utilized for mind wandering detection, as well as training and inference processes.

Data

We employ two different datasets in our work, one (Bosch & D'Mello, 2021) for within-dataset evaluation and the other (Lee et al., 2022) for cross-dataset evaluation. Table 1 provides a general overview of the datasets, and they are described as follows.

Lab Data by Bosch & D'Mello (2021)

This dataset is from a lab study, containing facial videos, recorded using a Logitech C270 webcam, of $N = 135$ university students from the U.S. reading a scientific text and their self-caught reports on mind wandering. Mind wandering instances and on-task moments of 10 s each were cut from the original videos (Fig. 1) (Bosch & D'Mello, 2021). The mind wandering instances are the time windows right before a self-report, with a 4 s buffer before the self-report, to ensure the exclusion of the key-press movement. The buffer length was validated in a pilot study. The on-task examples are 10-second clips taken from the time in-between, that do not include a page turn or fall into 30 s before a mind wandering self-report. The resulting dataset contains $N = 1031$ mind wandering instances and $N = 2406$ non-mind wandering instances. We use this dataset for training our models in both within- and cross-task evaluation scenarios as it contains more subjects and samples.

In-the-wild Data by Lee et al. (2022)

This open-source dataset available for research purposes contains facial videos of $N = 15$ university students in Korea and probe-based mind wandering reports. The

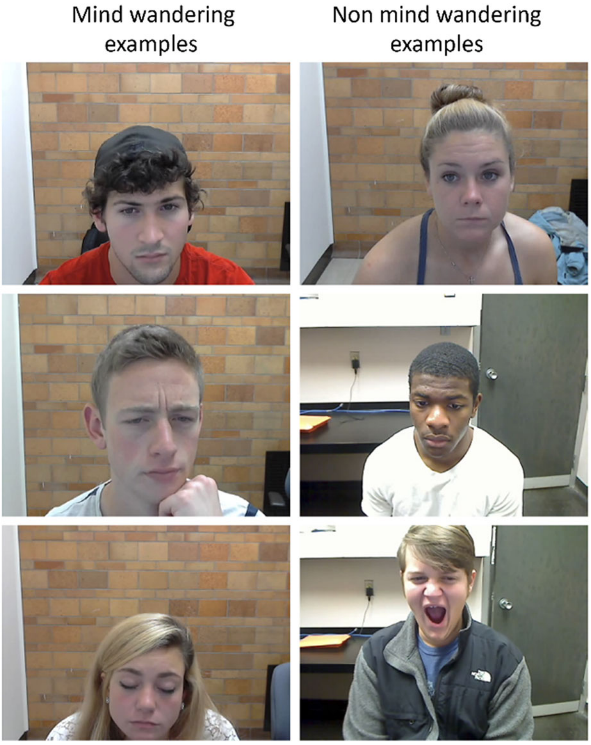

Fig. 1 Example images of mind wandering (left) and non mind wandering (right) instances in the lab data (Bosch & D'Mello, 2021)

participants watched a one-hour-long lecture video at home, were probed for mind wandering in 40-second intervals, and were filmed by their webcams. In contrast to previously employed datasets, the data was collected in the wild (i.e., at students'

Table 1 Dataset comparison

Specification	Datasets	
	Lab data	In-the-wild data
Study	Bosch & D'Mello (2021)	Lee et al. (2022)
Task	Reading scientific text	Watching lecture video
Setting	Laboratory	In the wild
Country	USA	Korea
Mind wandering self reports	Self-caught	Probe-caught
Participants	135	15
Total instances	3,437	1,220
Mind wandering	1,031	206
Non mind wandering	2,406	1,014
Video FPS	12.5	30

homes), which is an increasingly realistic learning setting for MOOCs and other online learning tools. In total, it contains $N = 205$ mind wandering and $N = 1009$ non-mind wandering instances with 30 FPS. Due to the smaller dataset size, we use this dataset solely for evaluation purposes to detect mind wandering in the wild.

Features

To extract deep learning-based facial expression features, we use a CNN with a ResNet50 (He et al., 2016) architecture pre-trained on the AffectNet dataset (Mollahosseini et al., 2017) containing 23,901 images classified as belonging to seven discrete facial expressions (neutral, happy, sad, surprise, fear, disgust, anger), as a feature extractor. The process is depicted exemplarily in Fig. 2. This model achieves an accuracy of 58% on the AffectNet validation set (see Sümer et al. (2021) for details on model training). To extract latent features from our video clips, we apply frame-wise face detection by employing RetinaFace (Deng et al., 2020) to our videos. In this step, we had to remove 74 instances, 31 of them mind wandering, because no face could be detected across all frames. We pre-process the resulting face images, aligning them based on five facial key points extracted by the face detector, then they are cropped to size 224 × 224 and normalized. We then insert them into the pre-trained ResNet50 model, from which the FER classification layer was removed. The extracted latent feature vector is 2048 digits long and can be fed into a downstream classifier. We provide insights into the most important image areas encoded in the feature vectors by applying the Explainable AI tool LIME (Ribeiro et al., 2016) to the feature extraction model.

To compare our latent deep learning features to more explicit features, similar to those used in previous research, we extract AUs, facial landmark locations, head location, pose and rotation, as well as face shape parameters. Additionally, we extract gaze direction vectors for both eyes and gaze angles, as well as 2D and 3D eye region landmarks, consisting of 55 landmark points for each. All features were extracted using the OpenFace toolkit (Baltrušaitis et al., 2016) from each video frame respectively.

Training and Inference

In our analysis approach for binary mind wandering classification, we aim to leverage pre-trained features from deep neural networks by employing transfer learning from the

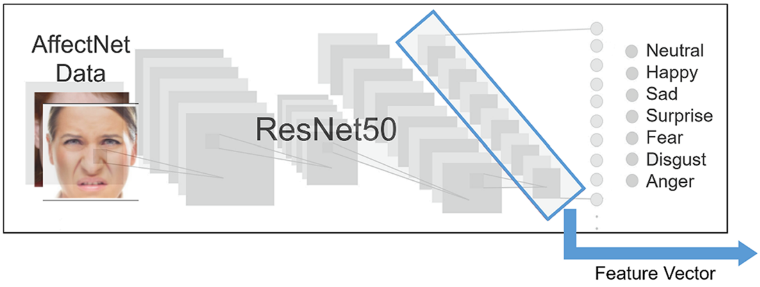

Fig. 2 Pre-trained Resnet50 as feature extractor

related task of facial expression recognition and compare this approach to employing explicit features as AUs and gaze vectors. Further, we examine whether these pre-trained features enable our model to generalize to a new in-the-wild dataset.

Handling Temporal Data

We aim to leverage temporal dynamics information which may be lost by aggregation over time in non-temporal models, by employing recurrent neural network models for supervised classification. In particular, we train long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and bidirectional long short-term memory (BiLSTM) (Baldi et al., 1999; Schuster & Paliwal, 1997) recurrent neural networks. For both models, we employ an architecture consisting of three recurrent layers with 100 neurons respectively, taking frame-wise extracted feature vectors with 125-time steps as input. A dense layer stacked on top, implementing a sigmoid activation function, outputs the binary mind wandering predictions.

The aforementioned self-reports serve as ground truth for our supervised machine learning approach. In order to compare our results in a within-task and within-context evaluation scenario, we employ the exact same person-independent 4-fold validation splits as in (Bosch & D'Mello, 2021). For each fold, a model is trained separately and results are averaged over all test folds. To emphasize the minority class mind wandering during training we employ class weighting. Furthermore, we employ early stopping of model training to avoid overfitting with a patience of 5 epochs.

To compare the performance of temporal models with non-temporal models, we additionally train SVMs (Cortes & Vapnik, 1995) with a radial basis function (RBF) kernel, which showed to be the best performing models in previous research (Bosch & D'Mello, 2021), as well as XGBoost (Chen & Guestrin, 2016) models and simple DNNs with one hidden layer (Fig. 3). To find the optimal parameter settings for each model, we used person-independent, nested 4-fold cross-validation to apply grid-search hyperparameter tuning. The best performing settings determined in inner 4-fold cross-validation were used for prediction in each outer fold. An overview of the parameters tested can be found in Table 7 in the appendix. To generate suitable input, we aggregate frame-wise extracted explicit features, computing the mean, median, minimum, maximum, and standard deviation values for each feature over the whole clip. For our latent features, we create statistical aggregations of the 2048-numbers long feature vector over all 125 frames. Since this procedure results in a large number of features with a lot of redundancy, we apply mutual information-feature selection, a univariate feature selection method based on the dependency between variables. Based on the training data, the 100 most meaningful features were selected in each fold. To account for the imbalanced data, we employ weighting or up-sampling of the training split using Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). We report the best performing combinations of balancing and hyper-parameters.

Fine-tuned CNN-LSTM

Employing temporal models allows us to not only employ the pre-trained AffectNet-CNN as a feature extractor but also train and fine-tune an end-to-end CNN-LSTM

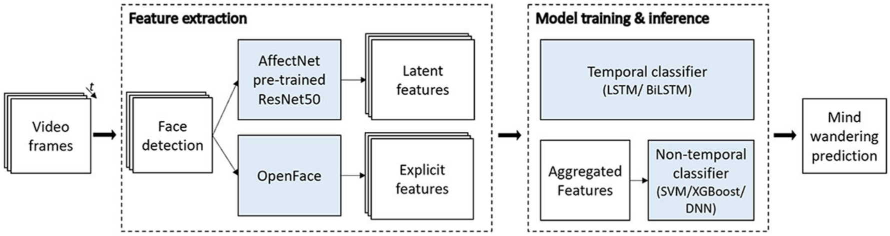
Video frames
t
Feature extraction
Face detection
AffectNet pre-trained ResNet50
Latent features
OpenFace
Explicit features
Model training & inference
Temporal classifier (LSTM/ BiLSTM)
Aggregated Features
Non-temporal classifier (SVM/XGBoost/ DNN)
Mind wandering prediction

Fig. 3 Mind wandering detection pipeline

model. This allows not only to train the temporal inference and classification parts of the model for our present problem of mind wandering detection but also to adapt the image feature extraction part of the model more precisely to the particularities of our problem. As a first step, the pre-trained ResNet50 without classification layer is included in the LSTM architecture to allow for video input and frame-wise feature extraction. Then, the previously described RNN architectures are stacked on top. In a first training run, the pre-trained weights in the CNN part of the model are frozen and only the top part of the model is allowed to adapt with a learning rate of 0.001. In a second training run, the last convolutional block of the ResNet50 model is set trainable and thus adapted to the mind wandering classification task with a smaller learning rate of 0.00001.

Cross-dataset Prediction

To examine the generalizability of the proposed approach over different settings, tasks, and target groups, we perform cross-dataset prediction employing our new approach. This means the aforementioned in-the-wild data is used to predict mind wandering instances. To this end, we use the video data from our lab reading task and train a model on the entire data. Using this model, we then predict mind wandering instances in the in-the-wild lecture viewing data from (Lee et al., 2022). We pre-process the in-the-wild dataset in the same way as the lab data. Based on the provided mind wandering probes, we cut 10-second windows before each probe and extracted all features described in "Features". For temporal models, we downsampled the 300 frames resulting from a higher recording frame rate to 125 frames to match the sequence length.

Evaluation Metrics

Due to the unbalanced nature of both our datasets with lab data having 30% and in-the-wild data having 25% mind wandering instances, reporting the accuracies of mind wandering prediction models could be misleading. A classifier always predicting non-mind wandering would achieve 70-75% accuracy without recognizing a single mind wandering instance. Also commonly used threshold metrics focusing on the minority class as F_1 scores are highly influenced by the skew in the data (Jeni et al., 2013). Such measures can be driven by high recall at low detection precision. For this reason, we report area under Precision-Recall curve (AUC-PR) values, which show the precision as function of the recall. This rank metric helps to balance precision and recall of the minority class.

Another measure that helps to evaluate the performance, especially when comparing performances of datasets with differing class distributions, is the improvement of the model above chance level, which is calculated as follows:

$$AboveChanceLevel = \frac{ActualPerformance - Chance}{PerfectPerformance - Chance}$$

To allow comparability to previous research, we additionally report the F$_1$ scores for the minority-class mind wandering, which is calculated as follows:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where the Precision (i.e., the proportion of correct mind wandering predictions of all mind wandering predictions) and Recall (i.e., share of correctly predicted mind wandering instances of all mind wandering instances) are defined as:

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}.$$

TP, FP, TN, and FN represent “true positive”, “false positive”, “true negative”, and “false negative”, respectively.

Results

In this section, we report within- and cross-dataset mind wandering detection results and investigate potential model biases by comparing our results across gender and explain the employed latent features.

Mind Wandering Detection

The results for within-task mind wandering detection, training, and predicting on the lab data, are depicted in Table 2. In the within-task prediction setting, models trained on latent features outperform the random baseline (i.e., $F_1 = 0.3$ (Bosch & D’Mello, 2021) and $AUC - PR = 0.3$) to a significant extent.

The results are comparable to the state-of-the-art on this data (Bosch & D’Mello, 2021), which achieved similar performance between F_1 scores of 0.414 and 0.478, mostly with hand-crafted features and SVMs. However, the observed F_1 scores are clearly driven by high recall values, while precision values are rather low. While a high recall means that most mind wandering instances are detected, simultaneously low precision also means that many instances are falsely classified as mind wandering. Therefore, we introduce the $AUC - PR$ score, which is a rank metric that also reflects the ratio of precision and recall. Thus, the above chance level values reported in the table are calculated on the basis of $AUC - PR$ scores.

Table 2 Results of mind-wandering detection for the within-lab-data prediction

Model	Feature Set	Method	AUC-PR	Above Chance-Level	F_1	Precision	Recall	ROC-AUC
SVM	Explicit features	OpenFace	0.288	-1.70%	0.391	0.286	0.636	0.482
	Latent features	AffectNet	0.365	9.30%	0.443	0.347	0.63	0.587
XG Boost	Explicit features	OpenFace	0.346	6.60%	0.332	0.348	0.321	0.543
	Latent features	AffectNet	0.372	10.30%	0.358	0.356	0.368	0.576
DNN	Explicit features	OpenFace	0.331	4.40%	0.367	0.333	0.439	0.507
	Latent features	AffectNet	**0.398**	**14.00%**	0.44	0.361	0.576	0.601
LSTM	Explicit features	OpenFace	0.323	3.23%	0.375	0.323	0.477	0.524
	Latent features	AffectNet	0.391	13.00%	**0.459**	0.362	0.636	**0.612**
BiLSTM	Explicit features	OpenFace	0.303	0.40%	0.335	0.288	0.462	0.483
	Latent features	AffectNet	0.383	11.90%	0.453	0.348	**0.658**	0.602
CNN-LSTM	Fine-tuned latent features	AffectNet	0.394	13.40%	0.41	**0.389**	0.445	0.605
CNN-BiLSTM	Fine-tuned latent features	AffectNet	0.384	12.00%	0.377	0.388	0.368	0.602

Random baseline lab data: F_1 = 0.3; Base rate = 0.3
The bold entries highlight best performing models

Table 3 Results of mind-wandering detection for cross-dataset prediction. Training on lab data, testing on in-the-wild data

Model	Feature Set	Method	AUC-PR	Above Chance-Level	F_1	Precision	Recall	ROC-AUC
SVM	Explicit features	OpenFace	0.214	5.30%	0.330	0.241	0.524	0.6
	Latent features	AffectNet	0.242	8.70%	0.301	0.261	0.355	0.596
XGBoost	Explicit features	OpenFace	0.18	1.20%	0.169	0.193	0.15	0.534
	Latent features	AffectNet	0.196	3.10%	0.193	0.194	0.192	0.574
DNN	Explicit features	OpenFace	0.218	5.80%	0.291	0.23	0.398	0.589
	Latent features	AffectNet	0.257	10.50%	0.296	**0.32**	0.276	0.646
LSTM	Explicit features	OpenFace	0.194	2.90%	0.283	0.217	0.408	0.551
	Latent features	AffectNet	0.252	9.90%	0.323	0.234	0.524	0.634
BiLSTM	Explicit features	OpenFace	0.168	-0.20%	0.251	0.16	0.583	0.47
	Latent features	AffectNet	0.261	11.00%	**0.352**	0.224	**0.825**	**0.658**
CNN-LSTM	Fine-tuned latent features	AffectNet	0.263	11.00%	0.238	0.308	0.195	0.652
CNN-BiLSTM	Fine-tuned latent features	AffectNet	**0.265**	**11.40%**	0.254	0.285	0.229	**0.658**

Random baseline in-the-wild data: F_1 = 0.25; Base rate = 0.17

The bold entries highlight best performing models

The results for cross-task mind wandering detection using in-the-wild data for evaluations are reported in Table 3. Similar to the within-task setting, our results outperform the random baseline (i.e., $F_1 = 0.25$ (Lee et al., 2022), $AUC - PR = 0.17$) despite the cross-task prediction and in-the-wild setting. State-of-the-art on this dataset (Lee et al., 2022) achieved F_1 scores between 0.25 and 0.36 using SVMs, XGBoost, and DNNs for mind wandering prediction in a within-dataset setting. We achieve a very comparable performance to the best performance of the state-of-the-art (Lee et al., 2022) despite cross-task evaluation, a culturally different target group, and an in-the-wild setting.

A comparison of performance along the employed feature sets over all classifiers and prediction scenarios, depicted in Fig. 4, indicates that overall the latent features allow a better mind wandering detection. Especially in cross-dataset prediction, the fine-tuning of the latent features in an end-to-end CNN-LSTM leads to further improvement. This could be due to the fact that more data, i.e., the complete lab data set, were available for the training, hence fine-tuning, than in the case of the within the prediction. When employing latent features, the use of temporal (i.e., LSTM, BiLSTM, CNN-LSTM, CNN-BiLSTM) models, allows for a small improvement of prediction performance compared to non-temporal classifiers (i.e., SVM, XGB, DNN) (Fig. 4, 2nd row), especially in the cross-dataset prediction.

Figure 5 shows the confusion matrices for within- and cross-dataset mind wandering prediction of the fine-tuned CNN-BiLSTM model, as well as precision-recall curves for both prediction scenarios. The confusion matrices reveal that the models are able to detect 37% and 23% of mind wandering instances in the two scenarios. Depending

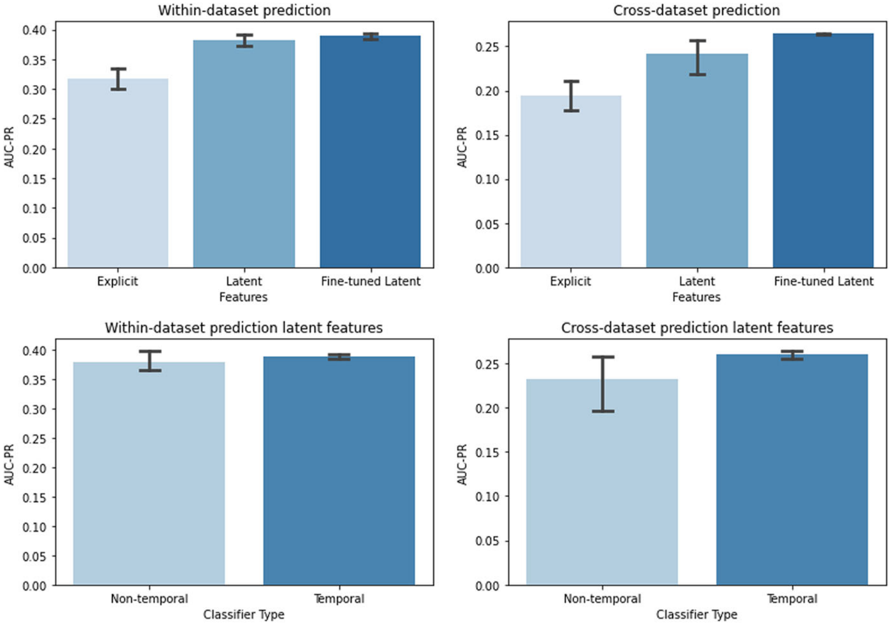

Fig. 4 Average Performance by Feature Sets and Classifier Type

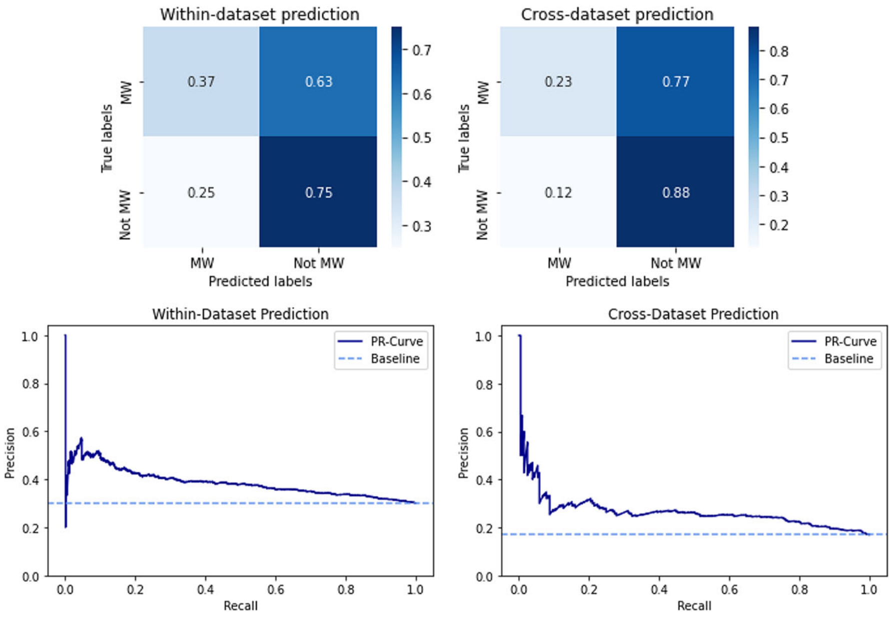

Fig. 5 Confusion Matrices and Precision-recall Curves for Within- and Cross-dataset Prediction with CNN-BiLSTM

on the detection use case high recall might be favored over precision, which can be achieved by lowering the prediction threshold.

When aggregating instance-by-instance predictions over time per participant, to assess performance over longer time windows, we obtain a Pearson's correlation coefficient of 0.578 ($P < 0.001$) between true and predicted mind wandering proportion per person in the within-dataset prediction. For cross-dataset predictions, we found a correlation of 0.48 ($P = 0.07$) on a person-level. However, when adapting the prediction threshold from 0.5 to 0.3 we achieve a correlation of 0.8 ($P < 0.001$). This finding underscores the potential to enhance mind wandering detection by fine-tuning the prediction threshold for specific use cases, particularly in cross-dataset predictions that exhibit varying overall proportions of mind wandering. Consequently, we conducted systematic experimentation to optimize thresholds for enhanced detection accuracy and applicability.

Threshold Optimization

In alignment with previous research (Stewart et al., 2017b; Faber et al., 2018) we employ threshold optimization to further improve the detection of mind wandering instances. Threshold optimization is a technique used to enhance the performance of a classification model by selecting the optimal cutoff value for distinguishing between the two classes. In binary classification, models often output a probability score that indicates the likelihood of an instance belonging to the positive class. The default threshold is typically set at 0.5. However, the default threshold of 0.5 may not always be the best choice for all datasets or objectives, especially in cases where there is class imbalance. Particularly in the cross-dataset scenario, where the two different datasets yielded varying mind wandering rates of 30% and 17%, threshold optimization emerges as a critical approach, enabling the adjustment of classification cutoffs to better reflect the disparate prevalence rates across datasets (Stewart et al., 2017b).

We systematically tested thresholds from 0.1 to 0.9 in increments of 0.1 for all employed models. Tables 4 and 5 show the results for optimization according to F_1 values for within- and cross-dataset predictions. Optimal thresholds are smaller than 0.5 for most classifiers. Those lower thresholds lead to notable improvements in F_1 values. Figure 6 shows evaluation metrics by prediction thresholds for CNN-BiLSTM models in both prediction scenarios. We can see that F_1 values drop at the 0.3 and 0.4 thresholds, which were identified as optimal for those models.

Comparison across Gender

To detect potential bias in the classification models and provide transparency with regard to the fairness of prediction across gender, we compare the detection performance of mind wandering by gender for the within- and cross-dataset predictions. We compare the results of all models trained on latent features for both prediction scenarios. The lab data employed for within-dataset predictions consists of 40.77% male and 59.23% female participants, with females reporting an overall higher mind wandering rate of 32.64% and males a rate of 26.41%.

The in-the-wild data, serving as an evaluation set for our cross-dataset predictions, contains 53.33% female and 46.67% male participants, reporting mind wandering in 18% and 16% of the mind wandering probes on average, respectively. The prediction results by gender are displayed in Table 6, with above chance values based on different base rates by gender and dataset. In general, mind wandering instances are predicted more accurately for females than for males. These results are most likely rooted in the imbalance of the underlying datasets, both regarding overall gender rates as well as gender-specific mind wandering rates, leading to overall fewer male mind wandering instances in the data. However, this difference becomes larger when employing temporal models, especially for within-dataset prediction scenarios. We can assume that this is due to the increased number of parameters in the temporal models, requiring large training data. While the overall performance increases, only females seem to benefit from improved detection, as there is more training data available.

Latent Feature Explanation

In order to gain deeper insights into the employed latent features, we apply the explainability algorithm LIME (Ribeiro et al., 2016) on our feature extraction CNN model, pre-trained on facial expression recognition task, and fine-tuned to our mind wandering detection task. LIME is an Explainable AI tool that helps to understand the decision-making of an algorithm, as it allows highlighting areas of interest in the image, the so-called super-pixels, that contribute positively or negatively to the model's prediction. This helps to get an intuition on why the model thinks this image belongs to a

Table 4 Results of threshold optimization for within-lab-data mind wandering prediction

Model	Threshold	Feature Set	Methods	AUC-PR	F1	Precision	Recall	ROC-AUC
SVM	0.5	Explicit features	OpenFace	0.288	0.391	0.286	0.636	0.482
	0.5	Latent features	AffectNet	0.365	0.443	0.347	0.63	0.587
XGBoost	0.1	Explicit features	OpenFace	0.346	0.434	0.313	0.703	0.543
	0.1	Latent features	AffectNet	0.372	0.456	0.334	0.717	0.576
DNN	0.1	Explicit features	OpenFace	0.331	0.461	0.3	**0.992**	0.507
	0.4	Latent features	AffectNet	**0.398**	0.467	0.338	0.752	0.601
LSTM	0.1	Explicit features	OpenFace	0.323	0.459	0.301	0.959	0.524
	0.4	Latent features	AffectNet	0.391	**0.48**	**0.345**	0.789	**0.612**
BiLSTM	0.1	Explicit features	OpenFace	0.303	0.459	0.302	0.954	0.483
	0.4	Latent features	AffectNet	0.383	0.478	0.34	0.806	0.602
CNN-LSTM	0.3	Fine-tuned latent features	AffectNet	0.394	0.472	0.334	0.806	0.605
CNN-BiLSTM	0.3	Fine-tuned latent features	AffectNet	0.384	0.475	0.334	0.823	0.602

Random baseline lab data: F_1 = 0.3; Base rate = 0.3
The bold entries highlight best performing models

Table 5 Results of threshold optimization for cross-dataset mind wandering prediction. Training on lab data, testing on in-the-wild data

Model	Threshold	Feature Set	Methods	AUC-PR	F1	Precision	Recall	ROC-AUC
SVM	0.5	Explicit features	OpenFace	0.214	0.324	0.231	0.539	0.6
	0.2	Latent features	AffectNet	0.242	0.329	0.225	0.611	0.596
XGBoost	0.1	Explicit features	OpenFace	0.18	0.286	0.188	0.597	0.534
	0.1	Latent features	AffectNet	0.196	0.297	0.182	0.793	0.574
DNN	0.4	Explicit features	OpenFace	0.218	0.316	0.216	0.592	0.589
	0.2	Latent features	AffectNet	0.257	0.349	0.248	0.586	0.646
LSTM	0.2	Explicit features	OpenFace	0.194	0.289	0.171	0.927	0.551
	0.5	Latent features	AffectNet	0.252	0.323	0.234	0.624	0.634
BiLSTM	0.1	Explicit features	OpenFace	0.168	0.3	0.177	**0.99**	0.47
	0.5	Latent features	AffectNet	0.261	0.352	0.224	0.825	**0.658**
CNN-LSTM	0.3	Fine-tuned latent features	AffectNet	0.263	**0.356**	0.235	0.737	0.652
CNN-BiLSTM	0.4	Fine-tuned latent features	AffectNet	**0.265**	0.353	**0.251**	0.595	**0.658**

Random baseline in-the-wild data: F_1 = 0.25; Base rate = 0.17
The bold entries highlight best performing models

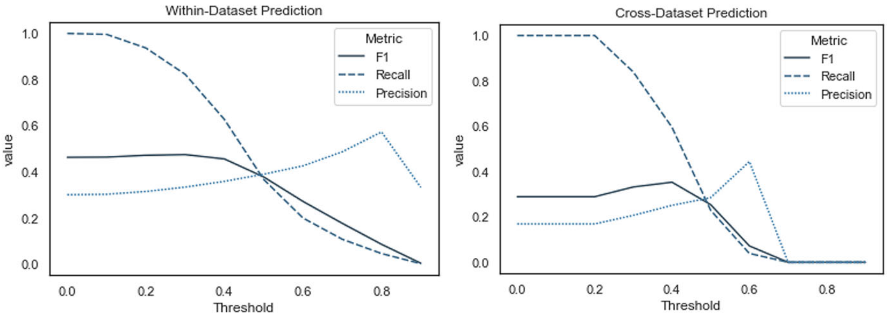

Fig. 6 Evaluation Metrics by Prediction Thresholds for Within- and Cross-dataset Prediction with CNN-BiLSTM

certain class and which part of a given image the decision is based on. It is important to take this step to uncover any unintended correlations that the classifier may have learned, such as those resulting from artifacts produced in the data collection process (Ribeiro et al., 2016).

In our study, these regions of interest let us draw conclusions about the information encoded in the latent features that are fed into the temporal mind wandering classifier. Therefore, it ensures the quality of feature extraction, by validating that the learned information aligns with established theoretical frameworks. Example images from the lab data including super-pixel boundaries and heatmaps, with dark blue encoding the most important areas, are provided in Fig. 7. The super-pixel boundaries include the 5 most important features positively contributing to the obtained prediction. The heatmaps depict super-pixels by importance, by coloring the most important areas

Table 6 Results of mind wandering detection using latent features by gender

Model	Gender	Within-dataset prediction		Cross-dataset prediction	
		AUC-PR	Above chance	AUC-PR	Above chance
SVM	Female	0.38	8%	0.283	12.6%
	Male	0.326	8.4%	0.189	3.5%
XGB	Female	0.365	5.7%	0.21	3.7%
	Male	0.301	5.0%	0.186	3.1%
DNN	Female	0.419	13.7%	0.351	20.9%
	Male	0.355	12.4%	0.186	3.1%
LSTM	Female	0.458	19.5%	0.335	18.9%
	Male	0.328	8.7%	0.196	4.3%
BiLSTM	Female	0.445	17.6%	0.335	18.9%
	Male	0.306	5.7%	0.186	0.7%
CNN-LSTM	Female	0.429	15.0%	0.29	13.4%
	Male	0.334	9.5%	0.207	5.6%
CNN-BiLSTM	Female	0.439	16.7%	0.333	18.7%
	Male	0.316	7.1%	0.199	4.6%

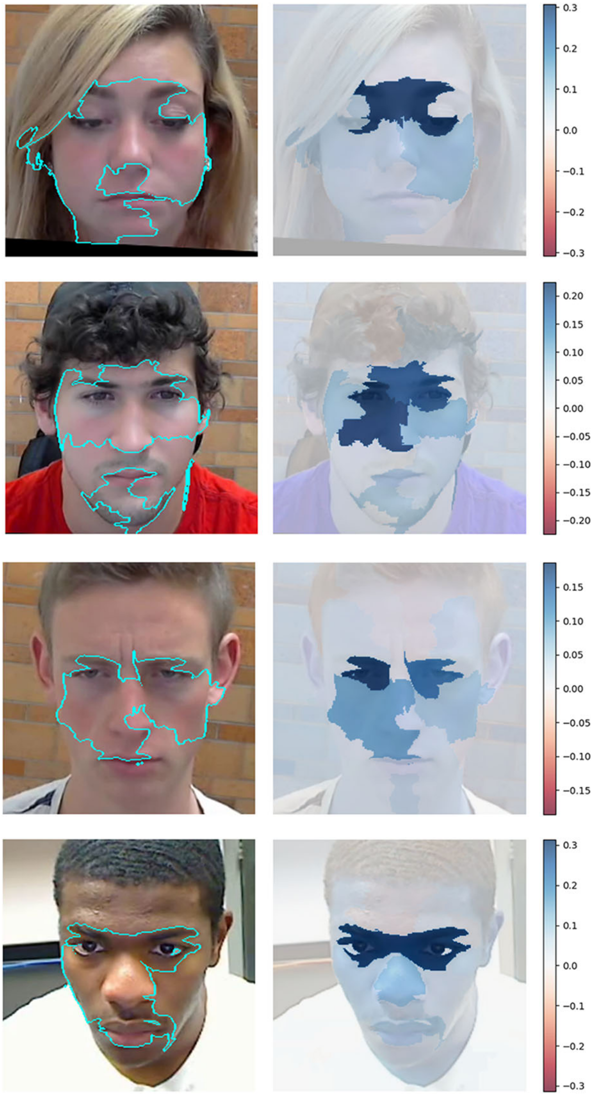

Fig. 7 Most important super-pixel boundaries (left) and super-pixel importance heatmaps (right) of pre-trained ResNet50 FER model, calculated with LIME for a true positive, true negative, false positive, and false negative samples in the final mind wandering classification (top to bottom)

in dark blue. We see that the CNN part of our classification model, employed for feature extraction, mainly relies on information from the eye and mouth areas of the participants, which are consequently inherently encoded in our feature vectors as well. We included example images based on our mind wandering prediction results including a true positive, true negative, false positive, and a false negative. These results are in line with previous studies suggesting facial textures and AUs are meaningful features for video-based mind wandering detection (Bosch & D’Mello, 2021).

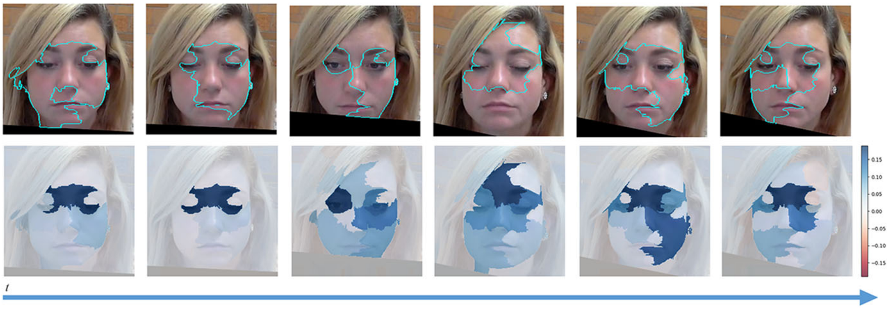

Fig. 8 Temporal dynamics of extracted latent feature super pixels. Example depiction of every 25th frame from one correctly predicted mind wandering instance

Since our analysis is based on video clips, the features shown are extracted from each frame individually and then transferred to the LSTM module as a time series. In this way, temporal dynamics can be displayed. To visualize this process, we have mapped the time course of the most important features for a correctly predicted mind wandering instance in Fig. 8. Every 25th frame from a 10 s clip is depicted. It is evident that the focused regions undergo minimal changes based on facial expression and eye opening, but the primary focus on the eye region remains consistent.

Discussion

In this section, we discuss our findings, also focusing on application scenarios and ethical issues.

Main Findings

We showed that features extracted from a pre-trained CNN on the AffectNet facial expression recognition dataset are informative for predicting mind wandering, even to a higher degree when they are utilized with temporal models that use frame-wise extracted features as input. The best results for within-dataset prediction were obtained using deep learning models such as DNN, LSTM, and BiLSTM, which allowed us to detect mind wandering $\approx$ 14% above the chance level. This indicates that the latent representations provided by a model based on the recognition of six basic facial expressions of emotion contain information at the same level for the task of mind wandering recognition as a set of explicit out-of-the-box features such as AUs, facial landmarks, and gaze vectors extracted by the OpenFace toolbox. Further, visualizations of areas of interest from the pre-trained FER model, fine-tuned on the mind wandering task showed that mainly eye and mouth areas were encoded by the model and, therefore, most likely encoded in the latent representation. these findings show that the features used represent meaningful information and can increase confidence in the latent representations. The results are comparable to those previously obtained by Bosch & D'Mello (2021) using a CART fusion of five explicit and hand-crafted feature groups on this data (F_1 = 0.478; ≈ 25% above chance level). By using transfer learning, the problem of insufficient data to train deep learning for this task can be overcome. Fine-tuning these features in an end-to-end model on the task at hand did slightly improve performance for the respective models but depends strongly on the available amount of data.

When we evaluated these models with a new dataset that differed in the task, setting, and target group, we showed that latent features generalized very well and achieved a prediction performance (≈ 11.4% above chance level) that was comparable to the best performance of a within-task prediction in the state-of-the-art (Lee et al., 2022). Also for the cross-dataset setting, the models on the latent features performed better overall than those on the OpenFace features. This indicates that the transfer-learned features generalize better than the explicit out-of-the-box features without manual feature engineering. This may be due to the fact that they are based purely on facial expression, which is less stimulus-specific than, for example, gaze. Beyond the only existing generalization study of video-based mind wandering recognition (Stewart et al., 2017b), which was based on AUs, we were able to show that latent facial features are also predictive beyond the lab in naturalistic settings, such as the home. These features were trained on an in-the-wild dataset, which might favor the transferability of the features to an in-the-wild setting.

Nevertheless, the generalization results are remarkable because not only the setting and task differ in our cross-dataset prediction, but also the culture of the target group. As mentioned in "Introduction", there is an ongoing debate about the universality of facial expressions of emotion (Russell, 1994; Ekman, 1994; Jack et al., 2012). One position taken is that the facial expressions of the six basic emotions relate to Western interpretations of these, but are not representative for instance, of East Asian culture (Jack et al., 2012). While often universality of facial expressions of the six basic emotions is assumed, several studies tackled cross-cultural FER (Dailey et al., 2010; Benitez-Garcia et al., 2017; Ali et al., 2020) and highlighted culture-specific differences in automated detection. The features used in this work are based on a model that was pre-trained on AffectNet data without accounting for culture-specific differences. Nevertheless, they generalize well from U.S. to Korean subjects in detecting mind wandering.

When comparing the prediction performances across gender, for both scenarios, we achieve better classification results for the female group. This difference could partly be the result of a combination of a greater proportion of women in both datasets, and the generally higher mind wandering rate for females in the training sample. However, gender differences are higher for temporal models with a higher number of parameters, in need of large training data. Another challenge is choosing the best prediction model. We demonstrated that adjusting the threshold for decision-making, depending on its application, can improve the effectiveness of these predictions, especially in cross-dataset prediction. However, focusing on F_1 scores can lead to an imbalance between recall and precision of the mind wandering class, therefore we focus on the rank metric AUC-PR. Depending on the application scenario, the need for optimization towards one or the other can be advantageous. While it is important to capture all mind wandering instances when used for research and testing of materials, it may be more useful to have high precision when using interventions not to disturb learners in attentive moments.

Applications

The ability of a machine learning model to generalize is of primary importance in terms of its potential applications. On the one hand, it enables finding the optimal design of learning materials and systems in a way that less mind wandering occurs. On the other hand, automated detection of mind wandering has the potential to help learners in online learning contexts such as ITSs and MOOCs to focus their attention on the task at hand, which often consists of tasks such as the ones we explored in this study: reading or video lecture viewing. This can be done through interventions such as providing feedback, suggesting re-watching or re-reading, asking intermediate questions, or adapting the presented content when a user loses focus. Such interventions can reduce user's mind wandering (D'Mello et al., 2017; Hutt et al., 2021; Mills et al., 2021) and therefore, based on the observed negative relationship between mind wandering episodes and learning success (Smallwood, 2011; Feng et al., 2013; Szpunar et al., 2013; Hollis & Was, 2016), improve learning outcomes. However, since all of these studies examine short-term effects, there is a need to validate the long-term impact of mind wandering interventions.

The automatic detection of learners' off-task thoughts presented here can be used in future research to provide learners with different types of interventions and feedback. An experimental examination with corresponding questionnaires and knowledge tests could examine whether these are suitable to help learners sustain their attention and whether they actually lead to improved learning success without being distracting or disruptive. While the efficacy of some interventions based on eye tracking has been investigated in both lab and classroom context (D'Mello et al., 2017; Hutt et al., 2021; Mills et al., 2021), a generalizable video-based recognition approach proposed in this study will allow future research to investigate the effectiveness of interventions outside the laboratory in naturalistic settings on a large scale. Furthermore, it will allow to deploy mind wandering detection modules in interfaces used in settings, in which eye tracking is not feasible.

While the current precision in predicting isolated occurrences of mind wandering might evoke uncertainties, our research shows that by consolidating predictions at the participant level correlations with the ground truth, ranging from moderate to high, can be achieved. This underscores that when aggregated over longer periods of time, analogous to the temporal scope utilized in interventions, models can yield improved outcomes. However, given the current performance of state-of-the-art video-based mind wandering prediction and the potential for false positive predictions, it is imperative to consider interventions that do not impede the learning process. Among the viable options, two seem most suitable: Non-interruptive interventions, like follow-up prompts to re-read or re-watch critical parts of learning materials, prioritize fostering self-regulated learning while minimizing disruptions and distractions. By seamlessly integrating these interventions into the learning environment, a more favorable and effective learning experience can be achieved. Furthermore, the implementation of thresholds, based on criteria such as prediction confidence or time durations, ensures the maintenance of intervention quality. This approach mitigates the risk of learner irritation or negative experiences resulting from inadequately controlled interventions. In doing so, even if video-based mind wandering detection is less performant than comparable tasks, such as emotion recognition, its use in online learning systems can still lead to an improvement in attention. It is crucial to determine the required level of precision for targeted interventions to attain their desired effect without impeding learning. In general, the goal should be to support learners in self-regulated learning (Järvelä et al., 2023) rather than to monitor or assess them. Therefore, it is important to consider user and data privacy in any potential application.

Ethical Considerations

While the application of video-based mind wandering detection holds promise for supporting learning in various applications, it also raises concerns regarding privacy, fairness, and inclusiveness. Since videos capture the identifiable faces of students during learning tasks, responsible handling of this data becomes imperative. While we work with raw videos, for model development, one of the advantages of our approaches is that for real world applications both explicit and latent features can be extracted on the fly in real time without the need of saving videos in the first place. A similar approach was employed in the classroom study of Bosch & D'Mello (2021) to preserve privacy. This aspect is very important as particularly vulnerable target groups such as school children (i.e., minors in dependent relationships), as those can also benefit from MOOCs and ITSs that utilize mind wandering detection. The HCI community has been addressing data privacy-related issues in different tasks such as face recognition (Erkin et al., 2009), gaze estimation (Bozkir et al., 2020), or even in the classroom context (Sümer et al., 2020) and similar approaches that preserve privacy should be utilized if one uses raw videos in an end-to-end fashion when mind wandering detection modules are deployed in real-world systems. Another convenient approach would also include utilizing federated learning systems (Li et al., 2020) where the training of machine learning models is carried out locally in user devices by keeping the sensitive data away from other parties and only sending the trained models to the computation party for aggregation. In brief, while the performance of mind wandering detection is an important aspect, privacy issues are of paramount importance for real-world use. In any case the process of collecting and analyzing any form of data should be transparent to the user with informed consent and clarity of the purposes the data will be used for (Nguyen et al., 2022b). Another important aspect of deploying automated mind wandering detection in educational settings is the inclusiveness in data and algorithms (Nguyen et al., 2022b). To ensure fairness and equality across user groups, it is crucial that these algorithms are trained and tested on diverse, unbiased data. This research underscores the significance of these considerations by aiming to develop a culturally generalizable algorithm and conduct a transparent analysis of gender fairness. While the results are promising with regard to generalization, especially the found gender differences stemming from unbalanced data, highlight the necessity for diverse and representative data that facilitate even better adherence to these principles in future research.

Limitations and Future Work

Despite its novelty, the presented work exhibits several limitations. First of all, we examined cross-setting predictions from the lab dataset to the in-the-wild dataset. This is partly due to the fact that lab-based data collection is more controlled, and therefore, allows for higher data quality. As reported in the original study (Lee et al., 2022) of the in-the-wild dataset, a large amount of data had to be excluded during data preparation due to insufficient data quality, for example, due to low luminance. From a practical standpoint, the generalizability from labeled lab data to in-the-wild data is a likely setting for future applications. Related to this point, the present in-the-wild dataset is rather small to train our proposed models on it. The need for a comparatively large labeled dataset to enable us to use the proposed temporal methods is a further limitation of this work. Currently, to our best knowledge, no larger in-the-wild mind wandering dataset is available. Despite the limited size of the in-the-wild dataset, this study examines the potential generalizability of state-of-the art mind wandering detection to diverse datasets. However, it is important to exercise caution when interpreting these findings due to the dataset's size. Future research should focus on collecting larger datasets to provide a more robust evaluation of the model's performance and enhance our understanding of its broader applicability. In general, further advances in automated detection of mind wandering should be evaluated based on their generalizability to in-the-wild settings to avoid the optimization of performance on single datasets at the expense of generalizability, as was recently discussed in the related field of affect detection (D'Mello & Booth, 2023).

While our cross-dataset prediction examines the models' ability to generalize two culturally different target groups, the sample of university and college students is somewhat homogeneous regarding other demographics, such as age. Future research should investigate, how well the model can be transferred to other target groups, for example, to school children. Furthermore, differences in predictive power by gender were revealed. A balanced sample by gender, as well as further investigation of gender differences in mind wandering, should be an important consideration for future studies to avoid bias.

In general, the presented results underscore the complexity of the recognition of mind wandering solely through facial video analysis. The approach presented in this paper, alongside previous approaches, can predict off-task thoughts above chance level, for dataset-specific class distributions, albeit without achieving exceptionally high prediction performances. While the achieved prediction accuracy may be considered moderate, and will potentially lead to false positive predictions when employed in educational systems, it is crucial to acknowledge the inherent difficulties associated with accurately identifying such a nuanced cognitive process. A prior study hinting at upper bounds for appearance-based detection showed that human observers only could identify mind wandering episodes to a similar extent (F_1= 0.406) as machine-learning based detection algorithms on the lab data (Bosch & D'Mello, 2022). These results strengthen the assumption that there is a limit to the accuracy with which the covert cognitive state of mind wandering can be detected based only on appearance. However, the required detection accuracy heavily depends on the desired application. Initial studies exploring interventions, like reiterating content, and asking questions, based on similar prediction performances show the potential success of those (D'Mello et al., 2017; Mills et al., 2021; Hutt et al., 2021). Future investigations should deploy the presented automated mind wandering detection approach to deliver interventions while assessing learning outcomes to determine the extent to which interventions improve attention and learning outcomes, whether frequent or incorrectly delivered mind wandering interventions hinder or interfere with learning and whether predictions at the current precision level are sufficient to be integrated into systems. As mentioned above, especially the use of non-intrusive interventions and the implementation of quality thresholds should be considered in the context of moderate prediction accuracies.

Additionally, the limited performance of these elaborate models can be attributed to the size constraints of the employed training dataset. Unfortunately, to our best knowledge, no other video-based mind wandering detection dataset is publicly available at this moment. To further advance the field and improve prediction outcomes, future research should place emphasis on collecting large-scale datasets, particularly in in-the-wild scenarios. By incorporating such diverse and extensive data, researchers can enhance the models' performance and pave the way for more accurate and robust mind wandering detection systems. A commendable approach was employed by Hutt et al. (2023), who gathered webcam eye tracking data both in university environments and via the platform Prolific, thereby ensuring the inclusion of diverse target groups and settings. Further potential strategies for enhancing prediction performance may include the implementation of personalization, such as the use of personalized prediction thresholds. This approach could partially compensate for the presumed interpersonal variations. Another way to improve prediction accuracy is to add other signals, like behavioural trace data or physiological signals. In this work, we focused on webcam videos due to the goal of scalability and low-threshold application in naturalistic settings. One potential for future research in this realm could be the further improvement of webcam-based eye tracking, since eye movements have been shown to be highly predictive for mind wandering (Hutt et al., 2023; Bixler & D'Mello, 2014; D'Mello et al., 2016; Faber et al., 2018; Mills et al., 2021).

Conclusion

In this work, we proposed a novel and generalizable approach for mind wandering detection utilizing facial features based on transfer learning from videos. Our results show the meaningfulness and transferability of those features with within- and cross-dataset prediction tasks on two challenging datasets. In particular, the results of the cross-dataset setting that differed with regard to the task, target group, and environment show the generalizability of our approach, which is key to the deployment of such models in intelligent learning systems to support learners to keep their attention on a given task.

Appendix A: Additional Tables

Table 7 Hyperparameter grids for employed non-temporal classification models

Model	Hyperparameter Grid
SVM	C (1, 10, 1000, 1000) Gamma (0.001, 0.01, 0.1)
DNN	Hidden layer size (100, 64, 32) Learning rate (0.001, 0.0001) Alpha (0.0001, 0.001, 0.005) Early stopping (True, False)
XGB	Gamma (0.5, 1, 2,5) Subsample (0.6, 0.8, 1) Column sample by tree (0.6, 1) Max depth (3, 5, 9) Learning rate (0.01, 0.1, 0.3)

Table 8 Results of mind-wandering detection for within- and cross-dataset prediction with feature level fusion

Prediction	Model	Method	F_1	Precision	Recall	AUC-PR	Above Chance-Level
Within-dataset	SVM	OpenFace, AffectNet	0.405	0.331	0.525	0.336	5.143%
	XG Boost	OpenFace, AffectNet	0.338	0.382	0.307	0.364	9.143%
	DNN	OpenFace, AffectNet	0.401	0.382	0.693	0.397	13.857%
	LSTM	OpenFace, AffectNet	**0.461**	0.334	**0.761**	0.373	10.429%
	BiLSTM	OpenFace, AffectNet	0.450	0.346	0.697	0.377	11.000%
Cross-dataset	SVM	OpenFace, AffectNet	0.299	0.218	0.473	0.207	4.5%
	XGBoost	OpenFace, AffectNet	0.202	0.198	0.207	0.199	3.5%
	DNN	OpenFace, AffectNet	0.311	0.238	0.448	0.220	6.0%
	LSTM	OpenFace, AffectNet	0.315	0.192	**0.864**	0.250	9.6%
	BiLSTM	OpenFace, AffectNet	0.335	0.214	0.772	0.229	7.1%

The bold entries highlight best performing models

Author Contributions All authors contributed to the study conception and design. Data analysis was performed by Babette Bühler. The first draft of the manuscript was written by Babette Bühler, Efe Bozkir and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. Babette Bühler is a doctoral candidate and supported by the LEAD Graduate School and Research Network, which is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II. This research was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - Project number 390727645. Sidney D'Mello was supported by the National Science Foundation (DRL 1920510). The opinions expressed are those of the authors and do not represent the views of the funding agencies.

Data Availability The Lab data by Bosch & D'Mello (Baldi et al. (1999)) is not publicly available. The In-the-wild data by Lee et al. (Lee et al. (2022)) is available for research purposes by contacting the authors at https://nmsl.kaist.ac.kr/projects/attention/.

Declarations

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Ali, G., Ali, A., Ali, F., Draz, U., Majeed, F., Yasin, S., & Haider, N. (2020). Artificial neural network based ensemble approach for multicultural facial expressions analysis. *IEEE Access, 8*, 134950–134963. https://doi.org/10.1109/ACCESS.2020.3009908

Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics, 15*(11), 937–946.

Baltrušaitis, T., Robinson, P., Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. *2016 ieee winter conference on applications of computer vision (wacv)* (p.1-10).

Benitez-Garcia, G., Nakamura, T., Kaneko, M. (2017). Analysis of in- and out-group differences between western and east-asian facial expression recognition. *2017 fifteenth iapr international conference on machine vision applications (mva)* (p.402-405).

Bixler, R., & D'Mello, S.K. (2014). Toward fully automated person-independent detection of mind wandering. *International conference on user modeling, adaptation, and personalization* (pp. 37–48).

Bixler, R., & D'Mello, S.K. (2021). Crossed eyes: Domain adaptation for gaze-based mind wandering models. *Acm symposium on eye tracking research and applications* (pp. 1–12).

Blanchard, N., Bixler, R., Joyce, T., D'Mello, S.K. (2014). Automated physiological-based detection of mind wandering during learning. Ş. Trăuşan-Matu, K.E. Boyer, M. Crosby, and K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 55–60). Cham: Springer.

Bonifacci, P., Viroli, C., Vassura, C., Colombini, E., Desideri, L. (2022). The relationship between mind wandering and reading comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 1–20, Bosch, N., & D'Mello, S. K. (2021). Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing, 12*(4), 974–988. https://doi.org/10.1109/TAFFC.2019.2908837

Bosch, N., & D'Mello, S. K. (2022). Can computers outperform humans in detecting user zone-outs? implications for intelligent interfaces. *ACM Transactions on Computer-Human Interaction, 29*(2), 1–33.

Bozkir, E., Ünal, A. B., Akgün, M., Kasneci, E., & Pfeifer, N. (2020). *Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework., 21*(1–21), 5. https://doi.org/10.1145/3379156.3391364

Brishtel, I., Khan, A. A., Schmidt, T., Dingler, T., Ishimaru, S., & Dengel, A. (2020). Mind wandering in a multimodal reading setting: Behavior analysis & automatic detection using eye-tracking and an eda sensor. *Sensors (Basel, Switzerland), 20*(9), 2546. https://doi.org/10.3390/s20092546

Caruso, M., & D'Mello, S. (2023). Do associations between mind wandering and learning from complex texts vary by assessment depth and time? *Lak23: 13th international learning analytics and knowledge conference* (pp. 230–239).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: ACM.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fmri reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106*(21), 8719–8724. https://doi.org/10.1073/pnas.0900234106

Conrad, C., & Newman, A. (2021). Measuring mind wandering during online lectures assessed with eeg. *Frontiers in Human Neuroscience, 15*, 697532.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J., & Cottrell, G. W. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion, 10*(6), 874.

Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5203–5212).

Dhawan, S. (2020). Online learning: A panacea in the time of covid-19 crisis. *Journal of educational technology systems, 49*(1), 5–22.

Dhindsa, K., Acai, A., Wagner, N., Bosynak, D., Kelly, S., Bhandari, M., & Sonnadara, R. R. (2019). Individualized pattern recognition for detecting mind wandering from eeg during live lectures. *PloS one, 14*(9), e0222276.

D'Mello, S. K., & Booth, B. M. (2023). Affect detection from wearables in the real wild: Fact, fantasy, or somewhere in between? *IEEE Intelligent Systems, 38*(1), 76–84.

D'Mello, S.K., Kopp, K., Bixler, R.E., Bosch, N. (2016). Attending to attention: Detecting and combating mind wandering during computerized reading. *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 1661–1669).

D'Mello, S.K., Mills, C., Bixler, R., Bosch, N. (2017). Zone out no more: Mitigating mind wandering during computerized reading. *International Educational Data Mining Society*.

D'Mello, S. K., & Mills, C. S. (2021). Mind wandering during reading: An interdisciplinary and integrative review of psychological, computing, and intervention research and theory. *Language and Linguistics Compass, 15*(4), e12412.

Dong, H. W., Mills, C., Knight, R. T., & Kam, J. W. (2021). Detection of mind wandering using eeg: Within and across individuals. *Plos one, 16*(5), e0251490.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. *Psychological Bulletin, 115*, 268–287.

Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I., Toft, T. (2009). Privacy-preserving face recognition. *Privacy enhancing technologies: 9th international symposium, pets 2009* (pp. 235–253).

Faber, M., Bixler, R., & D'Mello, S. K. (2018). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods, 50*(1), 134–150. https://doi.org/10.3758/s13428-017-0857-y

Faber, M., & D'Mello, S. K. (2018). How the stimulus influences mind wandering in semantically rich task contexts. *Cognitive Research: Principles and Implications, 3*(1), 35. https://doi.org/10.1186/s41235-018-0129-0

Faber, M., Krasich, K., Bixler, R. E., Brockmole, J. R., & D'Mello, S. K. (2020). The eye-mind wandering link: Identifying gaze indices of mind wandering across tasks. *Journal of Experimental Psychology: Human Perception and Performance, 46*(10), 1201–1221. https://doi.org/10.1037/xhp0000743

Feng, S., D'Mello, S. K., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review, 20*(3), 586–592. https://doi.org/10.3758/s13423-012-0367-y

Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *Quarterly Journal of Experimental Psychology, 66*(12), 2289–2294. https://doi.org/10.1080/17470218.2013.858170. (PMID: 24313285).

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 770–778). IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780.

Hollis, R. B., & Was, C. A. (2016). Mind wandering, control failures, and social media distractions in online learning. *Learning and Instruction, 42*, 104–112. https://doi.org/10.1016/j.learninstruc.2016.01.007

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., D'Mello, S.K. (2017). Gaze-based detection of mind wandering during lecture viewing. *International Educational Data Mining Society*.

Hutt, S., Krasich, K., Mills, C., Bosch, N., White, S., Brockmole, J. R., & D'Mello, S. K. (2019). Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction, 29*(4), 821–867. https://doi.org/10.1007/s11257-019-09228-5

Hutt, S., Krasich, K., R. Brockmole, J., K. D'Mello, S.K. (2021). Breaking out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–14).

Hutt, S., Wong, A., Papoutsaki, A., Baker, R.S., Gold, J.I., Mills, C. (2023). Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods*, 1–17,

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241–7244.

Jang, D., Yang, I., & Kim, S. (2020). Detecting mind-wandering from eye movement and oculomotor data during learning video lecture. *Education Sciences, 10*(3), 51. https://doi.org/10.3390/educsci10030051

Järvelä, S., Molenaar, I., & Nguyen, A. (2023). *Advancing srl research with artificial intelligence*. Elsevier.

Jeni, L.A., Cohn, J.F., De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. *2013 humaine association conference on affective computing and intelligent interaction* (p.245-251).

Jin, C. Y., Borst, J. P., & Van Vugt, M. K. (2019). Predicting task-general mind-wandering with eeg. *Cognitive, Affective, & Behavioral Neuroscience, 19*(4), 1059–1073.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive psychology, 8*(4), 441–480.

Kuvar, V., Kam, J.W., Hutt, S., Mills, C. (2023). Detecting when the mind wanders off task in real-time: An overview and systematic review. *Proceedings of the 25th international conference on multimodal interaction* (pp. 163–173).

Lee, T., Kim, D., Park, S., Kim, D., Lee, S- J. (2022). Predicting mind-wandering with facial videos in online lectures. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops* (pp. 2104–2113).

Lemay, D. J., Bazelais, P., & Doleck, T. (2021). Transition to online learning during the covid-19 pandemic. *Computers in Human Behavior Reports, 4*, 100130.

Levine, M.D. (1990). *Keeping a head in school.* Educators Publishing Service, Inc., 75 Moulton St., Cambridge, MA 02138-1104.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 1–1*,. https://doi.org/10.1109/TAFFC.2020.2981446

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Mills, C., Gregg, J., Bixler, R., & D'Mello, S. K. (2021). Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human-Computer Interaction, 36*(4), 306–332.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31.

Ngo, D., Nguyen, A., Dang, B., Ngo, H. (2024). Facial expression recognition for examining emotional regulation in synchronous online collaborative learning. *International Journal of Artificial Intelligence in Education*, 1–20,

Nguyen, A., Hong, Y., Dang, B., Nguyen, P.T.B. (2022). Emotional regulation in synchronous online collaborative learning: A facial expression recognition study. *ICIS 2022 Proceedings.*

Nguyen, A., Ngo, H.N., Hong, Y., Dang, B., Nguyen, B- P.T. (2022). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 1–21,

Pan, S. C., Sana, F., Schmitt, A. G., & Bjork, E. L. (2020). Pretesting reduces mind wandering and enhances learning during online lectures. *Journal of Applied Research in Memory and Cognition, 9*(4), 542–554.

Pham, P., & Wang, J. (2015). Attentivelearner: Improving mobile mooc learning via implicit heart rate tracking. C. Conati, N. Heffernan, A. Mitrovic, and M.F. Verdejo (Eds.), *Artificial intelligence in education* (pp. 367–376). Cham: Springer.

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: a theory-driven meta-analysis of attention regulation. *Psychological bulletin, 140*(6), 1411.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation, 29*(9), 2352–2449.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2012). Using ez reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye-mind link. *Psychological review, 119*(1), 155.

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 1135–1144).

Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology, 26*(2), 234–242. https://doi.org/10.1002/acp.1814

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin, 115*(1), 102.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15*(7), 319–326. https://doi.org/10.1016/j.tics.2011.05.006

Schubert, A- L., Frischkorn, G.T., Rummel, J. (2020). The validity of the online thought-probing procedure of mind wandering is not threatened by variations of probe rate and probe framing. *Psychological research, 84*(7), 1846–1856.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing, 45*(11), 2673–2681.

Smallwood, J. (2011). Mind-wandering while reading: Attentional decoupling, mindless reading and the cascade model of inattention. *Language and Linguistics Compass, 5*(2), 63–77. https://doi.org/10.1111/j.1749-818X.2010.00263.x

Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review, 14*(3), 527–533. https://doi.org/10.3758/BF03194102

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132*(6), 946–958.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology, 66*, 487–518. https://doi.org/10.1146/annurev-psych-010814-015331

Smilek, D., Carriere, J. S., & Cheyne, J. A. (2010). Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering. *Psychological science, 21*(6), 786–789.

Stewart, A., Bosch, N., Chen, H., Donnelly, P., D'Mello, S.K. (2017). Face forward: Detecting mind wandering from video during narrative film comprehension. *International conference on artificial intelligence in education* (pp. 359–370).

Stewart, A., Bosch, N., D'Mello, S.K. (2017). Generalizability of face-based mind wandering detection across task contexts. *International Educational Data Mining Society*.

Sümer, Ö., Gerjets, P., Trautwein, U., Kasneci, E. (2020). Automated anonymisation of visual and audio data in classroom studies. *The workshops of the thirty-fourth aaai conference on artificial intelligence.*

Sümer, Ö., Goldberg, P., D'Mello, S.K., Gerjets, P., Trautwein, U., Kasneci, E. (2021). Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*.

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: from the classroom to online learning. *Frontiers in Psychology, 4*, 495. https://doi.org/10.3389/fpsyg.2013.00495

Varao-Sousa, T. L., & Kingstone, A. (2019). Are mind wandering rates an artifact of the probe-caught method? using self-caught mind wandering in the classroom to test, and reject, this possibility. *Behavior Research Methods, 51*(1), 235–242. https://doi.org/10.3758/s13428-018-1073-0

Wong, A. Y., Smith, S. L., McGrath, C. A., Flynn, L. E., & Mills, C. (2022). Task-unrelated thought during educational activities: A meta-analysis of its occurrence and relationship with learning. *Contemporary Educational Psychology, 71*, 102098.

Zhang, H., Miller, K. F., Sun, X., & Cortina, K. S. (2020). Wandering eyes: Eye movements during mind wandering in video lectures. *Applied Cognitive Psychology, 34*(2), 449–464. https://doi.org/10.1002/acp.3632

Zhao, Y., Lofi, C., Hauff, C. (2017). Scalable mind-wandering detection for moocs: A webcam-based approach. É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, and M. Pérez-Sanagustín (Eds.), *Data driven approaches in digital education* (pp. 330–344). Cham: Springer International Publishing.

Authors and Affiliations

Babette Bühler[1] · **Efe Bozkir**[2,6] · **Patricia Goldberg**[1] · **Ömer Sümer**[3] · **Sidney D'Mello**[4] · **Peter Gerjets**[5] · **Ulrich Trautwein**[1] · **Enkelejda Kasneci**[6]

✉ Babette Bühler
babette.buehler@uni-tuebingen.de

Efe Bozkir
efe.bozkir@uni-tuebingen.de

Patricia Goldberg
patricia.goldberg@uni-tuebingen.de

Ömer Sümer
oemer.suemer@informatik.uni-augsburg.de

Sidney D'Mello
sidney.dmello@colorado.edu

Peter Gerjets
p.gerjets@iwm-tuebingen.de

Ulrich Trautwein
ulrich.trautwein@uni-tuebingen.de

Enkelejda Kasneci
enkelejda.kasneci@tum.de

[1] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, Tübingen 72072, Germany

[2] Human-Computer Interaction, University of Tübingen, Sand 14, Tübingen 72076, Germany

[3] University of Augsburg, Universitätsstraße 6a, Augsburg 86159, Germany

[4] Institute of Cognitive Science and the Department of Computer Science, University of Colorado, UCB 344/ 1905 Colorado Ave., Boulder 80309, Colorado, United States

[5] Leibniz-Institut für Wissensmedien, Schleichstraße 6, Tübingen 72076, Germany

[6] Human-Centered Technologies for Learning, Technical University Munich, Arcisstraße 21, Munich 80333, Germany