Detecting When the Mind Wanders Off Task in Real-time: An Overview and Systematic Review

Vishal Kuvar
kuvar001@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Julia W. Y. Kam
julia.kam@ucalgary.ca
University of Calgary
Calgary, Alberta, Canada

Stephen Hutt
stephen.hutt@du.edu
University of Denver
Denver, Colorado, USA

Caitlin Mills
cmills@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

ABSTRACT

Research on the ubiquity and consequences of task-unrelated thought (TUT; often used to operationalize mind wandering) in several domains recently sparked a surge in efforts to create "stealth measurements" of TUT using machine learning. Although these attempts have been successful, they have used widely varied algorithms, modalities, and performance metrics — making them difficult to compare and inform future work on best practices. We aim to synthesize these findings through a systematic review of 42 studies identified following PRISMA guidelines to answer two research questions: 1) are there any modalities that are better indicators of TUT than the rest; and 2) do multimodal models provide better results than unimodal models? We found that models built on gaze typically outperform other modalities and that multimodal models do not present a clear edge over their unimodal counterparts. Our review highlights the typical steps involved in model creation and the choices available in each step to guide future research, while also discussing the limitations of the current "state of the art" — namely the barriers to generalizability.

CCS CONCEPTS

• **Applied computing** → **Psychology**.

KEYWORDS

mind wandering detection; task-unrelated thought; machine learning; systematic review

ACM Reference Format:
Vishal Kuvar, Julia W. Y. Kam, Stephen Hutt, and Caitlin Mills. 2023. Detecting When the Mind Wanders Off Task in Real-time: An Overview and Systematic Review. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3577190.3614126

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614126

1 INTRODUCTION

At some point while you are reading this paper, your mind is likely to drift off to thinking about something else entirely – perhaps considering what you will eat for dinner tonight or a conversation you had earlier with a friend. Such shifts in attention away from our current task are often referred to as mind wandering [59], and are referred to here as *task-unrelated thought* (TUT; [72]). By most estimates TUT occurs 20-50% of the time while we are awake, regardless of our current "task" (e.g., conversations, reading, driving, etc. [21, 25, 48, 64]). Beyond its striking frequency, it has important functional correlates: on the positive end of the spectrum, it is associated with creativity ([1]), future planning [2], and can provide respite during boring periods ([66]). However, on the negative side, it is consistently linked to poorer task performance, including sustained attention [47, 55], comprehension [12, 68], driving [3, 80], memory [11, 79], and dysphoria [71].

Such negative associations are thought to be a result of diminished external processing [45, 67] – ultimately limiting our understanding of the information coming in [25, 68]. Recent work has thus focused on to creating detectors of TUT in order to build systems that can "respond" in real-time to assuage some of this negative impact. For example, a system that can detect when someone's mind wanders off task in real-time can provide support during learning (i.e. by helping fill gaps of any missed information), driving (i.e. by giving alerts), and other domains.

Despite increasing interest in this field, the best markers and approaches for TUT detection are still relatively opaque: what markers are the most helpful for prediction, in what tasks, and are multimodal models worth their cost? In the current research, we attempt to provide some clarity in this burgeoning area of research by taking stock of TUT detection through a systematic review. This research comes approximately a decade after the first TUT detection papers were published, with the aim of answering descriptive questions about the state of the field including: what has been done, what works best, and what trends may help guide future practices.

1.1 Background

1.1.1 Mind-body link. . The promise of TUT detection rests on the purported coupling between the mind and the body. TUT detection in particular, draws from past psychological research suggesting that behavioral and physiological signals (i.e., eye gaze, response

times, heart rate, neural signals, etc.) are reliably linked to TUT; that is, the brain and body show certain "markers" when the mind wanders [23]. For example, prior work suggests that our eye gaze (i.e., how long we fixate on certain parts of a screen [32]), electroencephalography signals (i.e., brain waves in various regions of our brain; [46]), and reaction times (i.e., how long it takes us to respond to other stimuli; [4, 62]) demonstrate marked changes when our minds go off-task.

These signals are arguably well suited to supervised machine learning with some markers unfolding at the millisecond level, beyond what a human may be able to identify without assistance. Indeed, research has shown that humans are typically unable to accurately predict whether someone is off-task based on the human eye [14].

1.1.2 Utility of TUT detection. The inability for humans to observe the TUT state as it arises in others presents a problem for research. As an example, consider how often you go off-task in the midst of a conversation. The ability to leave the "here and now" is evolutionarily beneficial, and, in most cases, we can safely assume the person we are talking to cannot tell our mind is elsewhere. But how do we measure this state for empirical purposes? Most research, including the work that initially established the link between TUT and behavioral/physiological signals, uses self-reports as "ground truth" for whether one's mind has wandered off-task [65, 77, 78]. These "gold standard" methods of assessments require participants to respond to thought probes that appear throughout the task (e.g., imagine a small box that pops up and asks, 'are you thinking about anything other than the task right now [yes/no]').

Such self-reports are not without some obvious drawbacks in terms of measurement; they necessarily interrupt the natural task flow while also requiring participants to make metacognitive judgments about their ongoing thoughts. They may also fail to capture many instances of TUT (for more discussion methods and issues, see [25]). The ability to accurately monitor TUT in real-time thus opens a range of possibilities for more controlled, ecological studies through stealth assessments of TUT.

Another key motivation for TUT detection is the ability to respond when the mind wanders in real-time. This is particularly compelling for domains where TUT seems to have the most detrimental impacts, such as education [79], navigation [33], medical diagnosis [70], and mental health [71]. For example, at least one study demonstrated that a real-time intervention when students experience TUT during reading could be effective for promoting deeper, longer-lasting learning gains – arguably by helping students "fill the gaps" in their understanding at the very moment they need support before they continue reading [58]. However, such one-off examples are currently at the proof-of-concept stage rather than a wide-reaching reality.

1.2 Novelty and Overview of Current Study

The TUT detection field is nascent compared to some other multimodal interaction/detection technologies. As an analogy, consider affect detection, where there is growing consensus about which features are predictive and rough baselines for what constitutes an "accurate" and state-of-the-art model [24, 49]. No such attempts at convergence have been made for TUT detection – a goal we tackle in the current work through the first systematic review. For example, what are the most popular and successful machine learning (ML) algorithms for TUT detection, and which feature modalities (gaze, neural, log files, etc.) are the best indicators of TUT? Our systematic literature review will explore the general trends in the field by considering the following varying elements of TUT detection: the tasks, feature modalities, as well as unimodal versus multimodal models. We will also provide takeaways from the review and forward-looking recommendations in the Discussion section, particularly by considering the diversity feature in the model building and the intended scale of the potential applications.

2 METHODS

2.1 Systematic Search

A systematic search was conducted between 6th March 2023 and 13th March 2023. Searches were conducted across the following resources: dblp Computer Science Technology (Schoss Dagstuhl), Education Resources Information Center (Institute of Education Sciences), PsychINFO (OVID), and Scopus (Academic Search Complete). In all four databases, titles, abstracts, and keywords were set as parameters to be searched across. Search terms that were used included "mind wandering detection," "real-time mind wandering detection," "predicting mind wandering," "detecting mind wandering," and "mind wandering detectors." Separate searches were conducted by once replacing the term mind wandering with "task-unrelated thought" and once with "off-task thought." In total, 15 terms each were searched in the four databases (see PRIMSA diagram in Figure 1).

2.2 Eligibility Criteria

Articles were required to meet the following criterion to be included in the review: First, the article included an empirical study (i.e., not a review, or the application of a previously developed detector). Second, TUT was a predicted (or classified) variable in analysis. Third, the article used supervised machine learning prediction with some cross-validation method (i.e., training data was not used in the testing set). The screening process was done in two phases. A researcher (one of the authors) went through the title, abstract, and keywords for each of the studies and eliminated papers that were clearly out of scope. This was followed by three researchers (authors) going through the selected articles to confirm inclusion.

2.3 Search Results and Data Extraction

The search yielded 276 articles in total. 115 were removed as they were duplicates. 119 were removed for not meeting the inclusion criteria above, leaving 42 unique articles that fit the inclusion criteria to be included in the systematic review.

Relevant information was then extracted from each paper, including: sample size and characteristics, task description and experimental set up, modality, best features, algorithms, performance metrics, and validation methods.

3 RESULTS

We first identified the commonalities across all papers in terms of the model building process. Six important elements were noted in

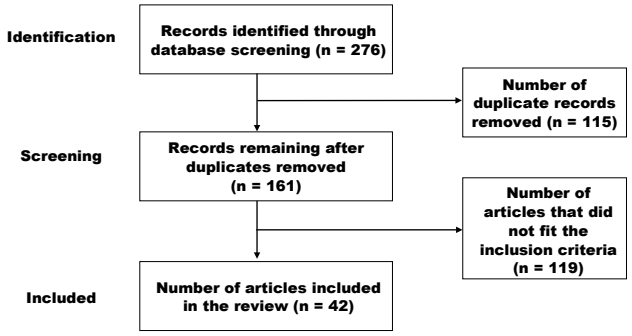


Figure 1: PRISMA diagram for systematic search.

each paper: 1) what task participants were doing, 2) how ground truth labels were collected, 3) what feature modality was collected, 4) the machine learning algorithm used, 5) how the model was validated, and 6) how it was evaluated. Figure 2 contains a quantitative summary of how frequent various options appeared in our review.

3.1 Performance Metrics

Reporting performance metrics is typically the final step in any machine learning pipeline. However, before discussing the general trends in terms of performance across different tasks and feature modalities, it is important to provide an overview of the common metrics used to evaluate TUT detectors. Note that some of the papers reported more than one evaluation metric, which is accounted for in the Ns below, and discussed in order from most common to least common.

Precision, Recall, F1 (N=22; Range: 0 to 1). A combination of Precision, Recall, and F1 were most commonly reported across studies. Precision is the proportion of true positives (TP) to the total number of instances labelled positive (TP + FP), whereas Recall is number of true positives divided by the number of instances that are positive (TP + FN). Finally, F1 score is the harmonic mean of precision and recall. However, F1 requires a "baseline score" for the predicted class of interest (in this case, the TUT class) to be compared against, which some of the reviewed papers did not report – presenting an area of improvement in future work.

Accuracy (N=18; Range: 0 to 1). Accuracy is usually the simplest metric to interpret as it simply represents the proportion of correctly identified points. However, accuracy is highly susceptible to misinterpretation for imbalanced data, as is the case for TUT (which generally occurred about 20-30% of the time in the studies reviewed). Predicting the majority class when TUT occurrs, for example, 20% of the time would lead to 80% accuracy, even if TUT was never predicted correctly. For this reason, it is difficult to interpret accuracy when chance accuracy for the TUT class is not reported, and we strongly encourage authors to avoid making claims about TUT detection using accuracy alone.

Kappa (N=13; Range: -1 to 1). Cohen's kappa has two major advantages. First, it considers imbalanced datasets and corrects for chance prediction. Second, it does not require a baseline score to be compared against. With these two advantages, kappa is sometimes interpreted as 'percentage increase over chance.' For example, a kappa score of zero would be interpreted as performing at chance levels and a score of 0.3 would be a 30% improvement over baseline.

AUROC (N=12; Range: 0 to 1). Area under curve or the receiver operating characteristic is another combination of precision and recall. It is calculated as the ratio of recall and the inverse of precision. With the range from zero to one, 0.5 represents the chance score and one represents perfect separation between classes.

Confusion matrix (N=10). Unlike the rest of the metrics on this list, a confusion matrix is not just one number, but rather a matrix that represents the distribution about classification correctness/incorrectness. In the case of binary TUT classification, "actual TUT" with "predicted TUT" are crossed to form a 2x2 table, allowing the true/false positives and negatives to be seen in a single table. Further, the matrix enables the calculation of other performance metrics and its inclusion in the results is highly encouraged. However, only 10 studies in our review reported the full matrix making it difficult to retroactively calculate additional metrics for comparison.

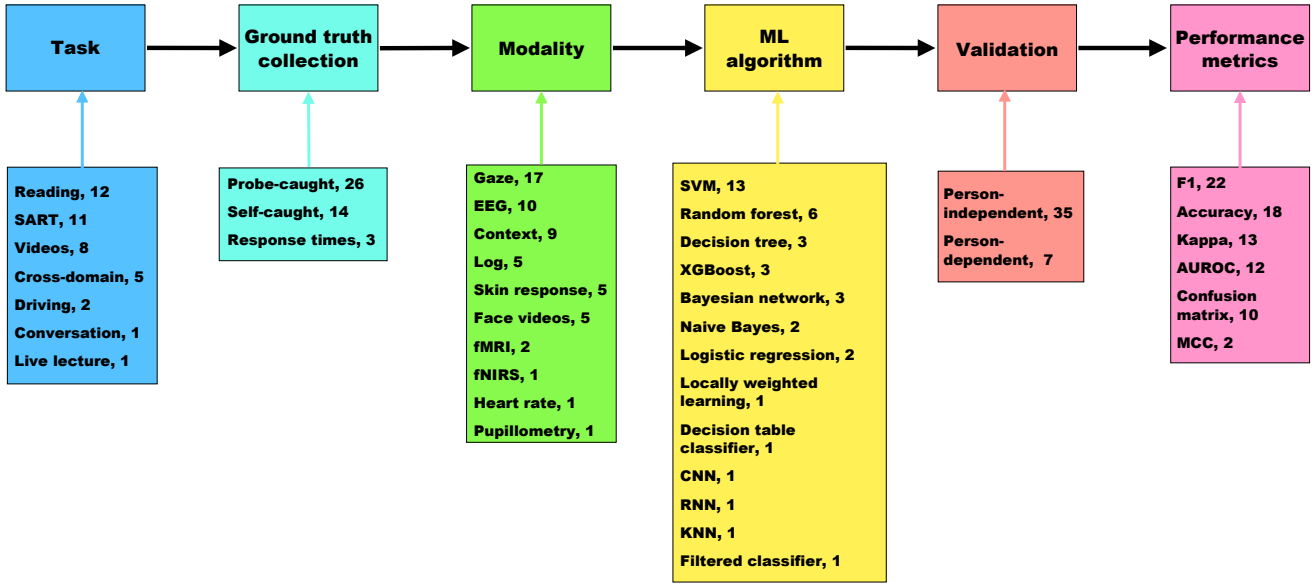
Matthew's Correlation Coefficient (N=2; Range -1 to 1). F1 uses false positive and false negative classifications in its calculations, but it leaves out the true negatives. This is not an issue in machine learning applications where the true negative is not of interest; however, it could be problematic in instances where on-task thought instances are also relevant. MCC takes into consideration the true negatives, with zero being chance and one being perfect classification. MCC, however, is the least used evaluation metric in our review.

3.2 Task

Now that we have covered the common metrics, we turn to a review and comparison of the tasks that are used in TUT detection. Below we make comparisons across tasks where possible (i.e., when similar evaluation metrics were used); this is easiest when using metrics like kappa and F1 scores given that they were both commonly used – however, we note that F1 should still be cautiously compared given the chance F1 scores will vary.

A "task" is defined as the primary activity a participant is engaged in (e.g., watching a video, reading, etc.). Broadly, the tasks observed in our review were categorized into seven categories: reading (N=12), sustained attention to response task (SART; N=11), watching videos (N=8), driving (N=2), computer-mediated conversations (N=1), live lecture (N=1), and cross-domain tasks (N=3). Out of these categories, reading and SART (a traditional cognitive psychology task) have been the most popular, possibly due to the high level of control these tasks provide to researchers. The top three tasks (reading, SART, and videos) together make up the majority of the TUT detection studies (N=31). Of the three, model performance is the lowest for videos, with best performance for reading, SART, and videos achieving kappas of 0.45, 0.318 and 0.23 respectively. The lower performance in videos might be attributable to the nature of the videos, as they are more dynamic and may provide a wider affordance landscape in terms of what participants may look at, think about, etc.

It is worth noting that even within a single category such as reading, there is still variability in the way that the task is presented. For example, while some studies that made use of the reading task



SART. sustained attention to response task; **EEG**. electroencephalogram; **fMRI**. functional magnetic resonance imaging; **fNIRS**. functional near-infrared spectroscopy; **SVM**. support vector machine; **CNN**. convolutional neural network; **RNN**. recurrent neural network; **KNN**. k-nearest neighbors; **AUROC**. area under the receiver operating characteristic curve; **MCC**. Matthew's correlation coeffcient

Figure 2: Descriptive Overview of Reviewed Studies Broken Down by the Six Elements in TUT detection.

showed the text one sentence at a time [57], others showed as much text as possible on each screen presented [9]. Similar variations exist for SART, which is a traditional cognitive psychology task measuring whether participants can accurately inhibit responses when certain characters are displayed on screen. In one study, participants were randomly shown letters of the alphabet and were asked to press a key for all except the target letter [19]. Another variation of the same SART task used auditory tones instead of letters [44]. There are also variations within videos; participants were either asked to watch films [73] or video lectures [39]. The Ns across these sub-categories were not large enough to form meaningful comparisons, but are worth noting as a possible "decision-point" for those extrapolating TUT detection methods.

Cross-task models. Most of the reviewed studies collected data from a single task, and a machine learning model was trained and tested from data within the same task. However, three studies in our sample trained cross-domain models – meaning they had multiple tasks as part of their research. The general idea behind this concept is to train a model on data from one task and test it on another (a paradigm very similar to transfer learning in most cases). Results about generalizability have been mixed. Models trained on SART were able to do well on visual search tasks and vice versa [44] showing potential generalizability between these two domains. Another study showed that although a model trained on data obtained while watching a film performed well on reading text, a model trained on text and tested on film performed at chance level [75]. In contrast, in the third study where models trained on reading illustrated text performed well when tested on a video lecture task, but the same trend was not seen when the tasks were switched [30].

3.3 Ground Truth Labels for Prediction

Ground truth is an inseparable part for supervised machine learning as it provides examples of positive and negative classes to the model. TUT research relies on two main techniques to gather ground truth-self-caught and probe-caught. The self-caught method (N=14) involves instructing the participant to monitor their thoughts and pressing a predetermined key to report TUT as soon as they notice it. This method has been commonly used and shown to be a reliable way to measure TUT [20], but it has two key drawbacks. First, it requires participants to be meta-aware of their thoughts throughout the task, effectively creating a dual-task situation. The second drawback is the lack of ground truth for the negative label (i.e., on-task thought).

The probe-caught technique (N=26) overcomes this problem by explicitly asking the participants to report TUT either using a binary or Likert scale. This in-situ method also does not require participants to monitor their own thoughts. Despite its popularity, a criticism of the probe-caught method is that it could impact the performance because it interrupts the natural flow of the task [65]. However, probes themselves do not appear to significantly affect performance when compared to the self-caught method [77].

Models trained on probe-caught data generally outperform models built on self-caught data. The highest kappa value achieved by the self-caught method was 0.363 [51] compared to a probe-caught value of 0.45 [8]. However, it's worth pointing out that since participants in the self-caught method only report instances of TUT, ground truth labels for "on-task" thought are retroactively created. Researchers typically use a windowed approach to generate the on-task labels [13, 56, 75]; when a TUT report is observed, a "window" of time is created to represent a TUT instance (e.g., 15 seconds

leading up to the report). In some studies, any remaining windows that fall outside of the TUT windows are then labeled as on-task thought. A potential issue is that participants can lose awareness of their thoughts and fail to report instances off-task thought, as TUT can occur without awareness [69]. This could ultimately result in instances of off-task thought being labelled as on-task, and might explain why the probe-caught method works better than self-caught.

Finally, a few studies (N=3) used reaction time and/or errors as a way to infer TUT occurrence. That is, if a participant makes an error or slow response, the authors assumed that it is because participants were not attending to the task [17, 38, 53].

3.4 Modality

3.4.1 Unimodal. In this section we describe the trends in performance associated with each modality when they are used in unimodal efforts. Synthesis of the analysis can be found in Table 1.

Gaze signals. (N=12) Gaze is the most popular modality for TUT detection. This popularity likely stems from a combination of the strong relationship between visual attention and TUT, as well as the increased affordability of eye tracking devices in recent years. Gaze-based features for TUT detection are broadly classified into two categories- *local* and *global* features [9, 56]. Local gaze features are dependent on the task content itself – e.g., the number of fixations that fall inside locations where participants are expected to look based on salience, importance, etc. Global features, in contrast, do not rely on the task and are arguably more simplistic in nature; they are calculated the same way regardless of what is happening in the task (e.g,. average fixation duration with no information about what the fixations were on). Although local features are helpful, particularly in more dynamic tasks where visual stimuli changes often, they are less transferable across tasks given the dependence on task content. Global features are inherently task-independent and thus may be better suited for generalizing across tasks.

Based on feature analyses reported in the papers, the gaze features that contributed "most" to the best models were dependent on the task that the participants performed. During reading, models trained on a combined set of global and local features outperformed models that were just trained on one set [28, 43]. Statistics of fixation duration (how long the eyes stay focused on something), like mean [31, 35], skew and kurtosis [8], were shown to be important across three different studies. Pupil diameter was also shown to be an important predictor of TUT [31, 35]. Unlike reading however, the best model trained for the video tasks only used global features. For example, raw number of saccades (periods of movement when their eyes do not focus on anything in particular) was a predictive feature that was negatively correlated with TUT [42, 83]. Gaze was also used to attempt domain general TUT detection across tasks such as reading, watching videos etc. [30]. Saccade amplitude and saccade velocity were the most predictive of TUT across domains.

Physiological signals. (N=12) Physiological and neural signals are the second most common unimodal feature choice. Across twelve studies, four different types of signals were used: EEG, fMRI, fNIRS, and Galvanic Skin response. As the majority of studies used EEG signals (n=9), we focus primarily on this signal. A primary difference between EEG-based TUT detection models exists in what data is given as input to the model. One study had success in detecting TUT with raw EEG signals [37]. Another study used the same raw signals as input to a pretrained image recognition model and used the activations in the penultimate layers as input [84]. Simply put, this allowed them to extract spatial relation between and within the signals to ultimately train a better model. However, it is important to note that these studies did not use person-independent validation and thus, their results might not replicate.

All other studies extracted specific features from the EEG signals. The most common features were event-related potentials [26, 44, 63, 76] and spectral power [22, 44, 54]. In addition to these linear features, a recent article examined the use of non-linear features, such as entropy, in their model [19]. Their feature analysis revealed three important features - two of which were linear and one non-linear in detecting TUT.

Unlike gaze-based signals, it is less apparent which EEG features consistently improve model performance. This could be attributed to the variety of metrics used for EEG signals whereas the set of gaze features were relatively similar across studies.

Face Videos. (N=4) A number of features can be extracted from face videos to create a TUT detection model. These include expressions, head movement, and subtle facial cue changes. Head pose was an important predictor in two studies [13, 75]. Models built on face videos have not performed as well as other modalities to date. For example, the biggest performance bump over chance was an F1 score of 0.407 (chance = 0.25). While models with these numbers exist in other modalities, usually, the performance over chance for these modalities is higher.

Log data. (N=3) Log data are often a "list" of actions/events performed by a user with corresponding timestamps, which can be used to extract behaviors. Other studies have used log data in tandem with other modalities (see section 3.4.2); however, log data itself can be an above chance predictor of TUT. Models trained on driving and keystroke features achieved a kappa score of 0.384 [5] and 0.363 [51], respectively. These studies suggest that simple log files, which are both scalable and do not require additional hardware, may be a viable way to detect TUT in real-time.

The features extracted from log files is quite variable, especially across tasks. During driving, distance from the car in front [5] and acceleration [82] were included in predictive models. During computer mediated chats, keystrokes were converted to word, sentence, and message level features that are indicative of TUT [51].

Does one feature modality perform better than the rest? Due to inconsistency over which metrics are reported in papers, comparing modalities head-to-head is difficult. However, we can make some conclusions based on how well models perform over the chance level. Models built on gaze seemed to perform better than other signals when metrics are compared against their respective baselines [29, 35, 38, 39]. A similar pattern emerges in the log data modality [5, 51, 82]. These speculations are not absolute as the efficacy of a modality also depends on the task.

3.4.2 Multimodal. Eleven studies used two or more modalities to predict TUT (see Table 2). Note that these studies were not reviewed above, as they are considered multi-modal. In addition to the feature modalities reviewed already, a final modality includes *context* features. Context-driven features tend to capture information not

Table 1: Search synthesis for unimodal models.

Modality Type	Modality	Task	Best Algorithm	Top Features	Model Validation	Performance Metric	Value
Log	Driving Behavior	Driving	Random Forest	-	PD	Accuracy	0.70 (chance = 0.61) [5]
			SVM	-	LOPO	Kappa	0.384 [82]
	Keystrokes	Conver-staion	Random Forest	Sentence length; word length	LOPO	Kappa	0.363 [51]
Gaze	Eye movements	Videos	Bayesian Network	Global features	LOPO	F1	0.47 (chance = 0.30) [39]
			Random Forest	Vergence	LOPO	F1	0.86 (chance = 0.6) [38]
			Naïve Bayes	Global features	LOPO	F1	0.405 (chance=NA) [83]
			Decision table classifier	Local features	LOPO	F1	0.59 (chance = 0.35) [56]
		Reading	Locally weighted learning	Global + Local features	LKPO	Kappa	0.23 [28]
			Logistic Regression	number of fixations; duration of fixations; pupil diameter	LOPO	AUROC	0.64 [31]
			Random Forest	mean fixation duration; mean saccade duration; pupil diameter	PD	AUROC	0.963 [35]
			SVM	Saccade duration; skew and kurtosis of fixation duration	LKPO	Kappa	0.45 [6]
			XGBoost	Global + Local features	LKPO	Kappa	0.15 [43]
		ITS	Bayesian Network	Global features	LKPO	F1	0.59 (chance = 0.24) [41]
		Cross-domain	SVM	Saccade velocity; Saccade amplitude	LOPO	Kappa	0.3 [30]
	Pupillometry	Reading	Random Forest	-	PD	Accuracy	0.84 (chance = NA) [81]
Physiological Signals	EEG	SART	SVM	Peak P3 amplitude	LOPO	MCC	0.206 [26]
			SVM	Mean amplitude; Mean energy	PD	Accuracy	0.89 (chance=NA) [76]
			Logistic Regression	-	PD	AUROC	0.68-0.95 [54]
			SVM	TP10	LOPO	Accuracy	0.756 (chance = NA)[63]
			Random Forest	-	LOPO	Kappa	0.318[19]
			CNN	-	PD	Accuracy	0.917 (chance = NA) [37]
			RNN	-	PD	MCC	0.84 [84]
		Live Lecture	SVM	theta=8; alpha=1; beta1=2 beta2=4	LOPO	F1	0.85 (chance = NA) [22]
		Cross-domain	SVM	Frontal alpha; Left occipital alpha	LOPO	Accuracy	0.60 (chance = 0.5) [44]
	fMRI	SART	-	DMN-FPCN	LOPO	Correlation r	0.11 [50]
	fNIRS	SART	XGBoost	-	LOPO	F1	0.732 [53]
	Physiology	SART	XGBoost	Mean GSR	LOPO	Kappa	0.294 [17]
Face Videos	Face Videos	Videos	SVM	-	LOPO	F1	0.30 (chance = NA) [74]
			SVM	Face vertical position	LOPO	F1	0.39 (chance = 0.29) [73]
		Cross-domain	Decision tree	jaw drop; lip tightener	LOPO	F1	0.407 (chance = 0.25) [75]
		Reading; ITS	SVM	upper body movement; head pose; facial textures	LOPO	F1	0.478; 0.414 (chance = NA) [13]

PD. person dependent; **LOPO**. leave-one-participant-out; **NA**. not available; **LKPO**. leave-k-persons-out; **AUROC**. area under the receiver operating characteristic curve; **SVM**. support vector machine; **ITS**. intelligent tutoring system; **MCC**. Matthew's correlation coefficient; **CNN**. convolutional neural network; **RNN**. recurrent neural network; **SART**. sustained attention to response task; **GSR**. Galvanic skin response

captured by another modality. Consider an example where gaze is used to build a TUT detector while reading [9]. Information about the text itself, such as difficulty level, cannot be captured using gaze. Context is the most common modality in the multimodal approach, with nine of the eleven studies utilizing some form of context. The modalities that have been used along with context are gaze (N=5), physiology (N=3), log data(N=2), face videos (N=1), and heart rate (N=1).

Gaze and context were the most consistently accurate modality combination. Two studies achieved kappa values very close to 0.3 [7, 9] and a third study reported an F1 score of 0.49 (chance=0.19; [42]). Physiology, on the other hand, is less consistent when added to multimodal models; although one study reported a kappa value of 0.41 [16], another study reported a kappa value of 0.19 [6] – one of the lowest kappas reported across all studies.

Is a multimodal approach better than unimodal? In most of these studies mentioned above, multimodal models outperformed the unimodal ones within the same study when they are directly compared [7, 9, 10, 16, 42, 52, 57]. However, multimodal models have not necessarily performed "better" as a rule in the broader literature, especially when comparing the results to the unimodal studies reviewed above. As an example, consider two instances when gaze and context features were combined in multimodal efforts to predict TUT during reading[7, 9]. Kappa values in these studies reached approximately 0.3, whereas another paper reported a kappa value of 0.45 when using gaze alone [8]. Such between-study differences in unimodal vs. multimodal model performance highlights the need to explore feature combinations in more detail. For example, performance of modalities might not be linearly additive, and may vary across tasks.

3.5 Algorithms

A total of 13 different algorithms, the set(s) of rules and procedures used for training, were found in this systematic review. Below we report the frequency with which each algorithm was used and which were identified as producing the "best" model (i.e. output produced after training using the algorithm).

Support Vector Machines (SVM) produced the best model in 13 studies. These models have been created using different tasks and modalities showing some extent of generalizability for SVMs. A similar trend is observed for random forests (N=6), though it was used about 50% less often compared to SVMs. In addition, decision trees, XGBoost, and Bayesian networks each produced the best model three times, with Naïve Bayes and logistic regression contributing two best models each. Finally, the following algorithms created the best models exactly once- locally weighted learning, decision table classifier, convolutional neural network, recurrent neural network, k-nearest neighbors, and filtered classifier. A direct contrast between these algorithms might not be entirely appropriate given that some algorithms, like XGBoost [18], are relatively newer compared to SVMs [15].

It is important to note that best models across studies that had used the same algorithm could have initialized them with a different set of hyperparameters. Reporting hyperparameters (or hyperparameter tuning) is usually out of scope for TUT detection papers, such that no trends can be observed.

3.6 Model Validation

Validation is a standard step in any machine learning pipeline. Here, we broadly classify model validation techniques as person-independent and person-dependent.

Person-independent (N=35). Person-independent validation is almost standard in TUT detection. This category of validation requires that data from a participant is exclusively included in the training or testing set, never both – giving a better idea of what the model would perform like on data it has never encountered before. Two variations of person-independent validations were used in studies included in this review. Leave-k-persons-out (LKPO; N=9) divides participants into groups such that each group has k participants, and each group is used as the "test" set exactly once. Leave-one-participant-out (LOPO; N=26), which is a special case of LKPO. It uses the same principle but creates as many groups as the number of participants in the study, where each person is the test set exactly once. Reported results typically reflect the average performance once each group/person has been included as the test set. The lower bound of performance for LOPO validation is an F1 score of 0.39, a 0.1 increase over chance [73], and the upper bound is a kappa of 0.384 [82]. On the other hand, performance for LKPO ranges from a kappa vale of 0.15 [43] to 0.45 [8].

Person-dependent (N=7). In this validation method, data from the same participant can be included in both training and testing set on the same run. Seven studies in this review used person-dependent validation. This method may be particularly appealing for instances when it makes sense to create a person specific classifier that will be used for the same person over time – perhaps in some learning contexts or with data that appear to have more individual signatures, such as EEG. However, it can also lead to overfitting (and reduced generalizability) when the training and test sets are not fully independent.

4 DISCUSSION

4.1 Summary of Main Findings

We reviewed a total of 42 articles to provide the first systematic review on TUT detection. Our goal was to assess the "state of the art" in this domain so that future efforts can have a resource for what features work best, in what situations, while also pointing out areas for standardization and improvement. We identified four key takeaways from our review: First, TUT detection is certainly feasible, though performance varies across tasks. Reading and SART were the most common tasks. This is likely due to the fact that TUT, which has exponentially grown in terms of research since 2006 [60], was primarily studied in highly controlled tasks until recently [40]. TUT detection was also more accurate in these tasks compared to those used less often. Second, gaze and EEG were the most prominent modalities used in unimodal approaches, whereas context was used the most in multimodal models. Third, multimodal models typically outperformed unimodal models when they were directly compared in the same studies, but unimodal efforts have been much more common to date. Fourth, SVMs commonly produced the "best" TUT detection models across modality or task.

Table 2: Search synthesis for multimodal models.

Modalities	Task	Best Algorithm	Top Features	Model Validation	Performance Metric	Value
Log; context	Reading	Decision Tree	Reading time; decoupling; number of characters	LOPO	Kappa	0.207 [57]
		Decision Tree	Minimum text pitch; disfluencies	PD	Accuracy	0.58 (chance = NA) [27]
Physiology; context	Reading	Filtered Classifier	-	LKPO	Kappa	0.22 [10]
Gaze; physiology; context	Reading	-	Fixation duration; skin temperature	LOPO	Kappa	0.19 [6]
		SVM	Pupil size; reading duration; saccade velocity	LOPO	Kappa	0.41 [16]
Gaze; context	ITS	NEAT	Skew of fixation duration; mean saccade duration	LKPO	F1	0.49 (chance = 0.19) [42]
	Reading	Naïve Bayes	Fixation duration; saccade length	LKPO	Kappa	0.28 [7]
		Bayes Net	Fixation length; number of words skipped	LOPO	Kappa	0.31 [9]
Gaze; face videos	Video	XGBoost		LKPO	AUROC	0.67 [52]
Heart rate; context	Videos	KNN		LOPO	Kappa	0.22 [61]
fMRI; EEG; Physiology	SART	SVM	Intra- and inter-network functional connectivity features	LOPO	F1	0.51 [34]

LOPO. leave-one-participant; **PD**. person-dependent; **LKPO**. leave-k-persons-out; **SVM**. support vector machine; **ITS**. intelligent tutoring system; **NEAT**. NeuroEvolution of Augmenting Topologies. **AUROC**. area under the receiver operating characteristic curve; **KNN**. k-nearest neighbors; **SART**. sustained attention to response task

4.2 Recommendations and Areas of Improvement

Sample diversity. Machine learning models, including TUT detectors, can be susceptible to bias. Barring a few studies in this review [17, 19, 37, 50, 53, 57, 76, 84], a majority of participants were high school or college students coming from mostly predominantly white serving institutions (an exception being [52]). The majority of the reviewed studies were also conducted in Western, Educated, Industrialized, Rich, and Democratized (W.E.I.R.D.) societies [36]. Thus, most of these models were trained on demographically skewed data and might not generalize to a broader population. To overcome this issue, diversifying samples should be a major priority in future work (e.g., race, ethnicity, neurodivergence, age, geography, etc.). This may allow for increased generalizability through more inclusive training data, or, at the very least, the ability to test for the existence of algorithmic bias (see [43]).

This point dovetails with a limitation of our own review; we only reviewed articles that were written in English, which may have resulted in the omission of relevant literature and ultimately biasing our conclusions about the diversity of samples used to date. Our systematic search was also done using the PsychInfo, dblp, Scopus, and ERIC databases, and we did not include all possible databases, such as the ACM Digital Library. At the same time, at least some of this risk is mitigated by the fact that several journals indexed in ACM are also cross-referenced in dblp, reducing the chance that a large number relevant articles were overlooked.

Generalizability. We suggest that researchers may want to be clear about the goals for detecting TUT, then decide what type of validation method is appropriate. As previously mentioned, person-dependent validation may result in higher performance metrics, yet be less likely to generalize to unseen individuals. To remedy this, the field should lean towards person-independent validation, unless there are valid reasons why person-specific models will be more helpful for an eventual application.

Reporting standards. We suggest authors report multiple performance metrics to facilitate comparison across studies (e.g., kappa, AUROC, and MCC). Further when accuracy F1 scores are used, they must be accompanied by baseline (chance) rates. At the same time, we acknowledge that the metrics listed in this review are neither exhaustive, nor flawless. Hence, whenever possible, confusion matrices could be reported. Confusion matrices allow researchers to calculate performance metrics retroactively, making comparisons across different studies much more feasible.

Ethical concerns and thoughtful applications. The points above also come with a huge caveat, and we need to be explicit about the goals for TUT detection. We do not endorse surveillance for the sake of surveillance, but rather when applications can be thoughtfully designed with (and for) target individuals in mind. For example, we do not think detection systems should be used to label individuals as "off-task" for supervisory reports, but instead can be used to assist at times when attentional lapses may mitigate understanding or even cause harm (i.e., in high stakes navigation), or to even encourage "off-task breaks" in other cases. Our review suggests that detecting an "off-task" mind is indeed possible. The next steps are to understand how, when, and why such models can be utilized for good.

ACKNOWLEDGMENTS

This work was supported by Templeton World Charity Foundation [grant number 30265]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the named funders.

REFERENCES

[1] Benjamin Baird, Jonathan Smallwood, Michael D. Mrazek, Julia W. Y. Kam, Michael S. Franklin, and Jonathan W. Schooler. 2012. Inspired by Distraction: Mind Wandering Facilitates Creative Incubation. *Psychological Science* 23, 10 (Oct. 2012), 1117–1122. https://doi.org/10.1177/0956797612446024 Publisher: SAGE Publications Inc.

[2] Benjamin Baird, Jonathan Smallwood, and Jonathan W. Schooler. 2011. Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition* 20, 4 (Dec. 2011), 1604–1611. https://doi.org/10.1016/j.concog.2011.08.007

[3] Carryl L. Baldwin, Daniel M. Roberts, Daniela Barragan, John D. Lee, Neil Lerner, and James S. Higgins. 2017. Detecting and Quantifying Mind Wandering during Simulated Driving. *Frontiers in Human Neuroscience* 11 (2017). https://doi.org/10.3389/fnhum.2017.00406 Publisher: Frontiers.

[4] Mikael Bastian and Jerome Sackur. 2013. Mind wandering at the fingertips: automatic parsing of subjective states based on response time variability. *Frontiers in Psychology* 4 (2013). https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00573

[5] John Beninger, Andrew Hamilton-Wright, Heather E. K. Walker, and Lana M. Trick. 2021. Machine learning techniques to identify mind-wandering and predict hazard response time in fully immersive driving simulation. *Soft Computing* 25, 2 (Jan. 2021), 1239–1247. https://doi.org/10.1007/s00500-020-05217-8

[6] Robert Bixler, Nathaniel Blanchard, Luke Garrison, and Sidney D'Mello. 2015. Automatic Detection of Mind Wandering During Reading Using Gaze and Physiology. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*. ACM Press, Seattle, Washington, USA, 299–306. https://doi.org/10.1145/2818346.2820742

[7] Robert Bixler and Sidney D'Mello. 2014. Toward Fully Automated Person-Independent Detection of Mind Wandering. In *User Modeling, Adaptation, and Personalization*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben (Eds.). Vol. 8538. Springer International Publishing, Cham, 37–48. https://doi.org/10.1007/978-3-319-08786-3_4 Series Title: Lecture Notes in Computer Science.

[8] Robert Bixler and Sidney D'Mello. 2015. Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. In *User Modeling, Adaptation and Personalization*, Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless (Eds.). Vol. 9146. Springer International Publishing, Cham, 31–43. https://doi.org/10.1007/978-3-319-20267-9_3 Series Title: Lecture Notes in Computer Science.

[9] Robert Bixler and Sidney D'Mello. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction* 26, 1 (March 2016), 33–68. https://doi.org/10.1007/s11257-015-9167-1

[10] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. 2014. Automated Physiological-Based Detection of Mind Wandering during Learning. In *Intelligent Tutoring Systems*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia (Eds.). Vol. 8474. Springer International Publishing, Cham, 55–60. https://doi.org/10.1007/978-3-319-07221-0_7 Series Title: Lecture Notes in Computer Science.

[11] Philippe Blondé, Jean-Charles Girardeau, Marco Sperduti, and Pascale Piolino. 2022. A wandering mind is a forgetful mind: A systematic review on the influence of mind wandering on episodic memory encoding. *Neuroscience & Biobehavioral Reviews* 132 (Jan. 2022), 774–792. https://doi.org/10.1016/j.neubiorev.2021.11.015

[12] Paola Bonifacci, Cinzia Viroli, Chiara Vassura, Elisa Colombini, and Lorenzo Desideri. 2023. The relationship between mind wandering and reading comprehension: A meta-analysis. *Psychonomic Bulletin & Review* 30, 1 (Feb. 2023), 40–59. https://doi.org/10.3758/s13423-022-02141-w

[13] Nigel Bosch and Sidney K. D'Mello. 2021. Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing* 12, 4 (Oct. 2021), 974–988. https://doi.org/10.1109/TAFFC.2019.2908837

[14] Nigel Bosch and Sidney K. D'Mello. 2022. Can Computers Outperform Humans in Detecting User Zone-Outs? Implications for Intelligent Interfaces. *ACM Transactions on Computer-Human Interaction* 29, 2 (Jan. 2022), 10:1–10:33. https://doi.org/10.1145/3481889

[15] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, Pittsburgh Pennsylvania USA, 144–152. https://doi.org/10.1145/130385.130401

[16] Iuliia Brishtel, Anam Ahmad Khan, Thomas Schmidt, Tilman Dingler, Shoya Ishimaru, and Andreas Dengel. 2020. Mind Wandering in a Multimodal Reading Setting: Behavior Analysis & Automatic Detection Using Eye-Tracking and an EDA Sensor. *Sensors* 20, 9 (April 2020), 2546. https://doi.org/10.3390/s20092546

[17] Sheng Chang, Yi-Ta Chen, and An-Yeu Andy Wu. 2021. Efficient Mind-wandering Detection System with GSR Signals on MM-SART Database. In *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, Coimbra, Portugal, 199–204. https://doi.org/10.1109/SiPS52927.2021.00043

[18] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 785–794. https://doi.org/10.1145/2939672.2939785

[19] Yi-Ta Chen, Hsing-Hao Lee, Ching-Yen Shih, Zih-Ling Chen, Win-Ken Beh, Su-Ling Yeh, and An-Yeu Wu. 2022. An Effective Entropy-Assisted Mind-Wandering Detection System Using EEG Signals of MM-SART Database. *IEEE Journal of Biomedical and Health Informatics* 26, 8 (Aug. 2022), 3649–3660. https://doi.org/10.1109/JBHI.2022.3187346

[20] Maria T. Chu, Elizabeth Marks, Cassandra L. Smith, and Paul Chadwick. 2023. Self-caught methodologies for measuring mind wandering with meta-awareness: A systematic review. *Consciousness and Cognition* 108 (Feb. 2023), 103463. https://doi.org/10.1016/j.concog.2022.103463

[21] Alexander Colby, Aaron Wong, Laura Allen, Andrew Kun, and Caitlin Mills. 2023. Perceived Group Identity Alters Task-Unrelated Thought and Attentional Divergence During Conversations. *Cognitive Science* 47, 1 (2023), e13236. https://doi.org/10.1111/cogs.13236 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.13236.

[22] Kiret Dhindsa, Anita Acai, Natalie Wagner, Dan Bosynak, Stephen Kelly, Mohit Bhandari, Brad Petrisor, and Ranil R. Sonnadara. 2019. Individualized pattern recognition for detecting mind wandering from EEG during live lectures. *PLOS ONE* 14, 9 (Sept. 2019), e0222276. https://doi.org/10.1371/journal.pone.0222276

[23] Mariana Rachel Dias da Silva, Marie Postma, and Myrthe Faber. 2022. Windows to the Mind: Neurophysiological Indicators of Mind Wandering Across Tasks. In *New Perspectives on Mind-Wandering*, Nadia Dario and Luca Tateo (Eds.). Springer International Publishing, Cham, 123–142. https://doi.org/10.1007/978-3-031-06955-0_7

[24] Sidney D'Mello and Jacqueline Kory. 2012. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, Santa Monica California USA, 31–38. https://doi.org/10.1145/2388676.2388686

[25] Sidney K. D'Mello and Caitlin S. Mills. 2021. Mind wandering during reading: An interdisciplinary and integrative review of psychological, computing, and intervention research and theory. *Language and Linguistics Compass* 15, 4 (2021), e12412. https://doi.org/10.1111/lnc3.12412 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12412.

[26] Henry W. Dong, Caitlin Mills, Robert T. Knight, and Julia W. Y. Kam. 2021. Detection of mind wandering using EEG: Within and across individuals. *PLOS ONE* 16, 5 (May 2021), e0251490. https://doi.org/10.1371/journal.pone.0251490

[27] Joanna Drummond and Diane Litman. 2010. In the Zone: Towards Detecting Student Zoning Out Using Supervised Machine Learning. In *Intelligent Tutoring Systems*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Vincent Aleven, Judy Kay, and Jack Mostow (Eds.). Vol. 6095. Springer Berlin Heidelberg, Berlin, Heidelberg, 306–308. https://doi.org/10.1007/978-3-642-13437-1_53 Series Title: Lecture Notes in Computer Science.

[28] Sidney D'Mello, Jonathan Cobian, and Matthew Hunter. 2013. Automatic Gaze-Based Detection of Mind Wandering during Reading. (2013), 2.

[29] Sidney K D'Mello, Caitlin Mills, Robert Bixler, and Nigel Bosch. 2017. Zone out no more: Mitigating mind wandering during computerized reading. (2017), 8.

[30] Robert E. Bixler and Sidney K. D'Mello. 2021. Crossed Eyes: Domain Adaptation for Gaze-Based Mind Wandering Models. In *ACM Symposium on Eye Tracking Research and Applications*. ACM, Virtual Event Germany, 1–12. https://doi.org/10.1145/3448017.3457386

[31] Myrthe Faber, Robert Bixler, and Sidney K. D'Mello. 2018. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods* 50, 1 (Feb. 2018), 134–150. https://doi.org/10.3758/s13428-017-0857-y

[32] Myrthe Faber, Kristina Krasich, Robert E. Bixler, James R. Brockmole, and Sidney K. D'Mello. 2020. The eye–mind wandering link: Identifying gaze indices of mind wandering across tasks. *Journal of Experimental Psychology: Human Perception and Performance* 46, 10 (July 2020), 1201. https://doi.org/10.1037/xhp0000743 Publisher: US: American Psychological Association.

[33] Cédric Gil-Jardiné, Mélanie Née, Emmanuel Lagarde, Jonathan Schooler, Benjamin Contrand, Ludivine Orriols, and Cédric Galera. 2017. The distracted mind on the wheel: Overall propensity to mind wandering is associated with road crash responsibility. *PLOS ONE* 12, 8 (Aug. 2017), e0181327. https://doi.org/10.1371/journal.pone.0181327 Publisher: Public Library of Science.

[34] Josephine M Groot, Nya M Boayue, Gábor Csifcsák, Wouter Boekel, René Huster, Birte U Forstmann, and Matthias Mittner. 2021. Probing the neural signature of

mind wandering with simultaneous fMRI-EEG and pupillometry. *NeuroImage* 224 (Jan. 2021), 117412. https://doi.org/10.1016/j.neuroimage.2020.117412

[35] Jacek Gwizdka. 2019. Exploring Eye-Tracking Data for Detection of Mind-Wandering on Web Tasks. In *Information Systems and Neuroscience*, Fred D. Davis, René Riedl, Jan vom Brocke, Pierre-Majorique Léger, and Adriane B. Randolph (Eds.). Vol. 29. Springer International Publishing, Cham, 47–55. https://doi.org/10.1007/978-3-030-01087-4_6 Series Title: Lecture Notes in Information Systems and Organisation.

[36] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *The Behavioral and Brain Sciences* 33, 2-3 (June 2010), 61–83; discussion 83–135. https://doi.org/10.1017/S0140525X0999152X

[37] Seyedroohollah Hosseini and Xuan Guo. 2019. Deep Convolutional Neural Network for Automated Detection of Mind Wandering using EEG Signals. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Niagara Falls NY USA, 314–319. https://doi.org/10.1145/3307339.3342176

[38] Kuo-Ting Huang, Christopher Ball, Jessica Francis, Rabindra Ratan, Josephine Boumis, and Joseph Fordham. 2019. Augmented Versus Virtual Reality in Education: An Exploratory Study Examining Science Knowledge Retention When Using Augmented Reality/Virtual Reality Mobile Applications. *Cyberpsychology, Behavior, and Social Networking* 22, 2 (Feb. 2019), 105–110. https://doi.org/10.1089/cyber.2018.0150

[39] Stephen Hutt, Jessica Hardey, Robert Bixler, Angela Stewart, Evan Risko, and Sidney K D'Mello. 2017. Gaze-based Detection of Mind Wandering during Lecture Viewing. (2017), 6.

[40] Stephen Hutt, Kristina Krasich, James R. Brockmole, and Sidney K. D'Mello. 2021. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445269

[41] Stephen Hutt, Caitlin Mills, Nigel Bosch, Kristina Krasich, James Brockmole, and Sidney D'Mello. 2017. "Out of the Fr-Eye-ing Pan": Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, Bratislava Slovakia, 94–103. https://doi.org/10.1145/3079628.3079669

[42] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J Donnelly, and Sidney K D'Mello. 2016. The Eyes Have It: Gaze-based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. (2016), 8.

[43] Stephen Hutt, Aaron Wong, Alexandra Papoutsaki, Ryan S. Baker, Joshua I. Gold, and Caitlin Mills. 2023. Webcam-based eye tracking to detect mind wandering and comprehension errors. *Behavior Research Methods* (Jan. 2023). https://doi.org/10.3758/s13428-022-02040-x

[44] Christina Yi Jin, Jelmer P. Borst, and Marieke K. van Vugt. 2019. Predicting task-general mind-wandering with EEG. *Cognitive, Affective, & Behavioral Neuroscience* 19, 4 (Aug. 2019), 1059–1073. https://doi.org/10.3758/s13415-019-00707-1

[45] Julia W. Y. Kam, Elizabeth Dao, James Farley, Kevin Fitzpatrick, Jonathan Smallwood, Jonathan W. Schooler, and Todd C. Handy. 2011. Slow Fluctuations in Attentional Control of Sensory Cortex. *Journal of Cognitive Neuroscience* 23, 2 (Feb. 2011), 460–470. https://doi.org/10.1162/jocn.2010.21443

[46] J. W. Y. Kam, T. Rahnuma, Y. E. Park, and C. M. Hart. 2022. Electrophysiological markers of mind wandering: A systematic review. *NeuroImage* 258 (Sept. 2022), 119372. https://doi.org/10.1016/j.neuroimage.2022.119372

[47] Michael J. Kane and Jennifer C. McVay. 2012. What Mind Wandering Reveals About Executive-Control Abilities and Failures. *Current Directions in Psychological Science* 21, 5 (Oct. 2012), 348–354. https://doi.org/10.1177/0963721412454875 Publisher: SAGE Publications Inc.

[48] Matthew A. Killingsworth and Daniel T. Gilbert. 2010. A Wandering Mind Is an Unhappy Mind. *Science* 330, 6006 (Nov. 2010), 932–932. https://doi.org/10.1126/science.1192439 Publisher: American Association for the Advancement of Science Section: Brevia.

[49] Dimitrios Kollias and Stefanos Zafeiriou. 2021. Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework. http://arxiv.org/abs/2103.15792 arXiv:2103.15792 [cs].

[50] Aaron Kucyi, Michael Esterman, James Capella, Allison Green, Mai Uchida, Joseph Biederman, John D. E. Gabrieli, Eve M. Valera, and Susan Whitfield-Gabrieli. 2021. Prediction of stimulus-independent and task-unrelated thought from functional brain networks. *Nature Communications* 12, 1 (Dec. 2021), 1793. https://doi.org/10.1038/s41467-021-22027-0

[51] Vishal Kuvar, Nathaniel Blanchard, Alexander Colby, Laura Allen, and Caitlin Mills. 2022. Automatically detecting task-unrelated thoughts during conversations using keystroke analysis. *User Modeling and User-Adapted Interaction* (Aug. 2022). https://doi.org/10.1007/s11257-022-09340-z

[52] Taeckyung Lee, Dain Kim, Sooyoung Park, Dongwhi Kim, and Sung-Ju Lee. 2022. Predicting Mind-Wandering with Facial Videos in Online Lectures. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, New Orleans, LA, USA, 2103–2112. https://doi.org/10.1109/CVPRW56347.2022.00228

[53] Ruixue Liu, Erin Walker, Leah Friedman, Catherine M. Arrington, and Erin T. Solovey. 2020. fNIRS-based classification of mind-wandering with personalized window selection for multimodal learning interfaces. *Journal on Multimodal User Interfaces* (June 2020). https://doi.org/10.1007/s12193-020-00325-z

[54] James S. P. Macdonald, Santosh Mathan, and Nick Yeung. 2011. Trial-by-Trial Variations in Subjective Attentional State are Reflected in Ongoing Prestimulus EEG Alpha Oscillations. *Frontiers in Psychology* 2 (2011). https://doi.org/10.3389/fpsyg.2011.00082

[55] Jennifer C. McVay and Michael J. Kane. 2009. Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 1 (Jan. 2009), 196–204. https://doi.org/10.1037/a0014104

[56] Caitlin Mills, Robert Bixler, Xinyi Wang, and Sidney K D'Mello. 2016. Automatic Gaze-Based Detection of Mind Wandering during Narrative Film Comprehension. (2016), 8.

[57] Caitlin Mills and Sidney D'Mello. 2015. Toward a Real-time (Day) Dreamcatcher: Sensor-Free Detection of Mind Wandering During Online Reading. (2015), 8.

[58] Caitlin Mills, Julie Gregg, Robert Bixler, and Sidney K. D'Mello. 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 (July 2021), 306–332. https://doi.org/10.1080/07370024.2020.1716762 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2020.1716762.

[59] Caitlin Mills, Quentin Raffaelli, Zachary C. Irving, Dylan Stan, and Kalina Christoff. 2018. Is an off-task mind a freely-moving mind? Examining the relationship between different dimensions of thought. *Consciousness and Cognition* 58 (Feb. 2018), 20–33. https://doi.org/10.1016/j.concog.2017.10.003

[60] Samuel Murray, Kristina Krasich, Jonathan W. Schooler, and Paul Seli. 2020. What's in a Task? Complications in the Study of the Task-Unrelated-Thought Variety of Mind Wandering. *Perspectives on Psychological Science* 15, 3 (May 2020), 572–588. https://doi.org/10.1177/1745691619897966 Publisher: SAGE Publications Inc.

[61] Phuong Pham and Jingtao Wang. 2015. AttentiveLearner: Improving Mobile MOOC Learning via Implicit Heart Rate Tracking. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo (Eds.). Vol. 9112. Springer International Publishing, Cham, 367–376. https://doi.org/10.1007/978-3-319-19773-9_37 Series Title: Lecture Notes in Computer Science.

[62] Jason G. Randall, Frederick L. Oswald, and Margaret E. Beier. 2014. Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin* 140, 6 (Nov. 2014), 1411–1431. https://doi.org/10.1037/a0037428

[63] Chutimon Rungsilp, K. Piromsopa, A. Viriyopase, and K. U-Yen. 2021. MIND-WANDERING DETECTION MODEL WITH ELECTROENCEPHALOGRAM. In *18th International Conference Cognition and Exploratory Learning in Digital Age 2021*. IADIS Press. https://doi.org/10.33965/celda2021_202108L030

[64] Paul Seli, Roger E. Beaty, James Allan Cheyne, Daniel Smilek, Jonathan Oakman, and Daniel L. Schacter. 2018. How pervasive is mind wandering, really?,. *Consciousness and Cognition* 66 (Nov. 2018), 74–78. https://doi.org/10.1016/j.concog.2018.10.002

[65] Paul Seli, Jonathan Carriere, Merrick Levene, and Dan Smilek. 2013. How few and far between? Examining the effects of probe rate on self-reported mind wandering. *Frontiers in Psychology* 4 (2013). https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00430

[66] Joshua Shepherd. 2019. Why does the mind wander? *Neuroscience of Consciousness* 2019, 1 (Jan. 2019), niz014. https://doi.org/10.1093/nc/niz014

[67] Jonathan Smallwood, Emily Beach, Jonathan W. Schooler, and Todd C. Handy. 2008. Going AWOL in the brain: mind wandering reduces cortical analysis of external events. *Journal of Cognitive Neuroscience* 20, 3 (March 2008), 458–469. https://doi.org/10.1162/jocn.2008.20037

[68] Jonathan Smallwood, Daniel J. Fishman, and Jonathan W. Schooler. 2007. Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin & Review* 14, 2 (April 2007), 230–236. https://doi.org/10.3758/BF03194057

[69] Jonathan Smallwood, Merrill McSpadden, and Jonathan W. Schooler. 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 (June 2007), 527–533. https://doi.org/10.3758/BF03194102

[70] Jonathan Smallwood, Michael D Mrazek, and Jonathan W Schooler. 2011. Medicine for the wandering mind: mind wandering in medical practice. *Medical Education* 45, 11 (2011), 1072–1080. https://doi.org/10.1111/j.1365-2923.2011.04074.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2923.2011.04074.x.

[71] Jonathan Smallwood, Rory C. O'Connor, Megan V. Sudbery, and Marc Obonsawin. 2007. Mind-wandering and dysphoria. *Cognition & Emotion* 21, 4 (June 2007), 816–842. https://doi.org/10.1080/02699930600911531

[72] Jonathan Smallwood and Jonathan W. Schooler. 2015. The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annual Review of Psychology* 66, 1 (Jan. 2015), 487–518. https://doi.org/10.1146/annurev-psych-

010814-015331

[73] Angela Stewart, Nigel Bosch, Huili Chen, Patrick Donnelly, and Sidney D'Mello. 2017. Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension. In *Artificial Intelligence in Education*, Elisabeth André, Ryan Baker, Xiangen Hu, Ma. Mercedes T. Rodrigo, and Benedict du Boulay (Eds.). Vol. 10331. Springer International Publishing, Cham, 359–370. https://doi.org/10.1007/978-3-319-61425-0_30 Series Title: Lecture Notes in Computer Science.

[74] Angela Stewart, Nigel Bosch, Huili Chen, Patrick J. Donnelly, and Sidney K. D'Mello. 2016. Where's Your Mind At?: Video-Based Mind Wandering Detection During Film Viewing. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM, Halifax Nova Scotia Canada, 295–296. https://doi.org/10.1145/2930238.2930266

[75] Angela Stewart, Nigel Bosch, and Sidney K D'Mello. 2017. Generalizability of Face-Based Mind Wandering Detection Across Task Contexts. (2017), 8.

[76] Nadia Jebin Tasika, Mohammad Hasibul Haque, Mohsena Begum Rimo, Mohtasim Al Haque, Salwa Alam, Tasmi Tamanna, Md Anisur Rahman, and Mohammad Zavid Parvez. 2020. A Framework for Mind Wandering Detection using EEG Signals. In *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, Dhaka, Bangladesh, 1474–1477. https://doi.org/10.1109/TENSYMP50017.2020.9230790

[77] Yana Weinstein. 2018. Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods* 50, 2 (April 2018), 642–661. https://doi.org/10.3758/s13428-017-0891-9

[78] Yana Weinstein, Henry J. De Lima, and Tim van der Zee. 2018. Are you mind-wandering, or is your mind on task? The effect of probe framing on mind-wandering reports. *Psychonomic Bulletin & Review* 25, 2 (April 2018), 754–760. https://doi.org/10.3758/s13423-017-1322-8

[79] Aaron Y. Wong, Shelby L. Smith, Catherine A. McGrath, Lauren E. Flynn, and Caitlin Mills. 2022. Task-unrelated thought during educational activities: A meta-analysis of its occurrence and relationship with learning. *Contemporary Educational Psychology* 71 (Oct. 2022), 102098. https://doi.org/10.1016/j.cedpsych.2022.102098

[80] Matthew R. Yanko and Thomas M. Spalek. 2014. Driving With the Wandering Mind: The Effect That Mind-Wandering Has on Driving Performance. *Human Factors* 56, 2 (March 2014), 260–269. https://doi.org/10.1177/0018720813495280 Publisher: SAGE Publications Inc.

[81] Baojie Yuan, Yetong Han, Jialu Dai, Yongpan Zou, Ye Liu, and Kaishun Wu. 2020. I am Smartglasses, and I Can Assist Your Reading. In *Algorithms and Architectures for Parallel Processing*, Meikang Qiu (Ed.). Vol. 12453. Springer International Publishing, Cham, 383–397. https://doi.org/10.1007/978-3-030-60239-0_26 Series Title: Lecture Notes in Computer Science.

[82] Yuyu Zhang and Takatsune Kumada. 2018. Automatic detection of mind wandering in a simulated driving task with behavioral measures. *PLOS ONE* 13, 11 (Nov. 2018), e0207092. https://doi.org/10.1371/journal.pone.0207092

[83] Yue Zhao, Christoph Lofi, and Claudia Hauff. 2017. Scalable Mind-Wandering Detection for MOOCs: A Webcam-Based Approach. In *Data Driven Approaches in Digital Education*, Élise Lavoué, Hendrik Drachsler, Katrien Verbert, Julien Broisin, and Mar Pérez-Sanagustín (Eds.). Vol. 10474. Springer International Publishing, Cham, 330–344. https://doi.org/10.1007/978-3-319-66610-5_24 Series Title: Lecture Notes in Computer Science.

[84] Lillian Zhu, Feng Zhu, and Jodi Price. 2022. TopographyNET: a deep learning model for EEG-based mind wandering detection. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Northbrook Illinois, 1–10. https://doi.org/10.1145/3535508.3545533