Karimah and Hasegawa
Smart Learning Environments (2022) 9:31
https://doi.org/10.1186/s40561-022-00212-y

Smart Learning Environments

REVIEW

Open Access

Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods

Shofiyati Nur Karimah[1*†] and Shinobu Hasegawa[2†]

†Shofiyati Nur Karimah and Shinobu Hasegawa contributed equally to this work.

*Correspondence: sn-karimah@jaist.ac.jp

[1] Graduate School of Advanced Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan
[2] The Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan

Abstract

Background: Recognizing learners' engagement during learning processes is important for providing personalized pedagogical support and preventing dropouts. As learning processes shift from traditional offline classrooms to distance learning, methods for automatically identifying engagement levels should be developed.

Objective: This article aims to present a literature review of recent developments in automatic engagement estimation, including engagement definitions, datasets, and machine learning-based methods for automation estimation. The information, figures, and tables presented in this review aim at providing new researchers with insight on automatic engagement estimation to enhance smart learning with automatic engagement recognition methods.

Methods: A literature search was carried out using Scopus, Mendeley references, the IEEE Xplore digital library, and ScienceDirect following the four phases of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA): identification, screening, eligibility, and inclusion. The selected studies included research articles published between 2010 and 2022 that focused on three research questions (RQs) related to the engagement definitions, datasets, and methods used in the literature. The article selection excluded books, magazines, news articles, and posters.

Results: Forty-seven articles were selected to address the RQs and discuss engagement definitions, datasets, and methods. First, we introduce a clear taxonomy that defines engagement according to different types and the components used to measure it. Guided by this taxonomy, we reviewed the engagement types defined in the selected articles, with emotional engagement (n = 40; 65.57%) measured by affective cues appearing most often (n = 38; 57.58%). Then, we reviewed engagement and engagement-related datasets in the literature, with most studies assessing engagement with external observations (n = 20; 43.48%) and self-reported measures (n = 9; 19.57%). Finally, we summarized machine learning (ML)-based methods, including deep learning, used in the literature.

Conclusions: This review examines engagement definitions, datasets and ML-based methods from forty-seven selected articles. A taxonomy and three tables are presented

to address three RQs and provide researchers in this field with guidance on enhancing smart learning with automatic engagement recognition. However, several key challenges remain, including cognitive and personalized engagement and ML issues that may affect real-world implementations.

Keywords: Engagement estimation, Engagement definitions, Engagement datasets, Engagement methods

Introduction

Recognizing learners' engagement can provide insight for enhancing learner-educator, learner-learning material, and learner-learner interactions (Sumer et al., 2021). Learner engagement has been found to be positively correlated with academic achievement (Lei et al., 2018), and higher engagement levels lead to better learning outcomes (Ponitz et al., 2009). A good engagement state is associated with curiosity, interest, optimism, and passion, which enhances motivation to continue learning and pursue achievement (Fredricks et al., 2004). Therefore, engagement is an essential component in the learning process that may reduce dropout rates, increase productivity and learning, and provide insights for improving course content and lecture plans (Alexander et al., 1997; Fredricks et al., 2004).

Research on engagement can be considered from two perspectives (Leite et al., 2015). Robot/computers/agents can be viewed as supports for increasing human engagement (Yun et al., 2012; Hall et al., 2014; Rich et al., 2010; Sanghvi et al., 2011) or as tools for automatically estimating human engagement (McDuff et al., 2012; Whitehill et al., 2014; Nakano and Ishii 2010; Castellano et al., 2009; Minsu et al., 2013; Castellano et al., 2012). In this article, we mainly focus on the second perspective.

Moreover, engagement estimation methods can be divided into three categories: manual, semiautomatic, and automatic (Dewan et al., 2019). In traditional offline classrooms, educators can recognize engagement levels directly or use manual observation checklists and rating scales. In contrast, in distance learning settings, learner engagement is more difficult to estimate due to limitations with learner-educator interactions. Therefore, a smart learning setting that allows automatic engagement estimation is one of potential solutions for addressing this limitation.

Recent improvements in computational hardware and software that support classic machine learning and deep neural networks have led to promising research on automatic engagement estimation (Gudi et al., 2015; Chaouachi et al., 2010). In particular, with the outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), work in this field has increased considerably (Abdellaoui et al., 2020). Automatic engagement has become an important topic in several fields, such as human interaction research, including human-human interactions (HHIs) (Chatterjee et al., 2021), human-robot interactions (HRIs) (Yun et al., 2012; Rudovic et al., 2018b; Yue et al., 2019; Yun et al., 2020), human-computer interactions (HCIs) (Dubovi 2022; Monkaresi et al., 2017; Psaltis et al., 2018), and embodied conversational agents (ECAs) (AlZoubi et al., 2012). Furthermore, classroom applications are critical for improving smart education (Zaletelj and Košir 2017; Sumer et al., 2021; Ashwin and Guddeti 2020b).

Several automatic engagement estimation methods have been proposed in recent years. Among them, computer vision-based techniques are the most popular because

nonverbal behaviours (including head motion, eye gaze, and body pose) play key roles in determining engagement levels (Ben-Youssef et al., 2019). In addition, computer vision-based approaches offer unobtrusive assessments, similar to classroom situations where teachers observe learners without interrupting their activities. These methods are also cost-effective and usable in the near term (D'Mello et al., 2017). Therefore, in this article, we conduct a systematic review on computer vision-based automatic engagement estimation methods that utilize appearance-based cues (such as videos or images).

Some physiological information-based methods that have received considerable attention in automatic engagement estimation research are also discussed. The development of cost-effective biosignal hardware, such as electroencephalogram (EEG), electrocardiogram (ECG), facial electromyogram (fEMG), and galvanic skin response (GSR), has provided simple and easy-to-use solutions (Alarcão and Fonseca 2019). Moreover, physiological signals support personalized analyses, which is pertinent for learners with special needs, such as those with autism (Rudovic et al., 2018b).

In this review, we aim to provide new researchers and educators in smart education and distance learning settings with an overview of the primary requirements and methods used to develop automatic engagement estimation methods, particularly in education/learning settings. The definitions, datasets, and methods are summarized in benchmark tables to provide an accessible overview of the systematic frameworks.

The research questions that guided this review were defined as follows:

- RQ1: How should the type of engagement to be measured be defined?
- RQ2: What datasets are suitable for developing automatic engagement estimation methods?
- RQ3: What automatic engagement estimation methods have been developed in the literature?

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) to select articles for this review. In the literature, we first reviewed the definition of engagement to help new researchers understand what type of engagement they want to focus on. Understanding the type of engagement being focused on is important before engagement levels are measured to determine which engagement cues, datasets and methods should be used. The widely used datasets and machine learning-based methodologies for automatic engagement estimation are then examined.

The remainder of this article is organized as follows. The procedure for selecting the articles for this review is explained in "Review method" section. "Results and discussions" section presents the key finding based on the RQs. Finally, the conclusions, including the contributions, limitations, and future directions, are summarized in "Conclusion" section.

Review method

The systematic review methodology employed in this study was adopted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) model (Page et al., 2021). The review structure was guided from PRISMA 2020 Checklist

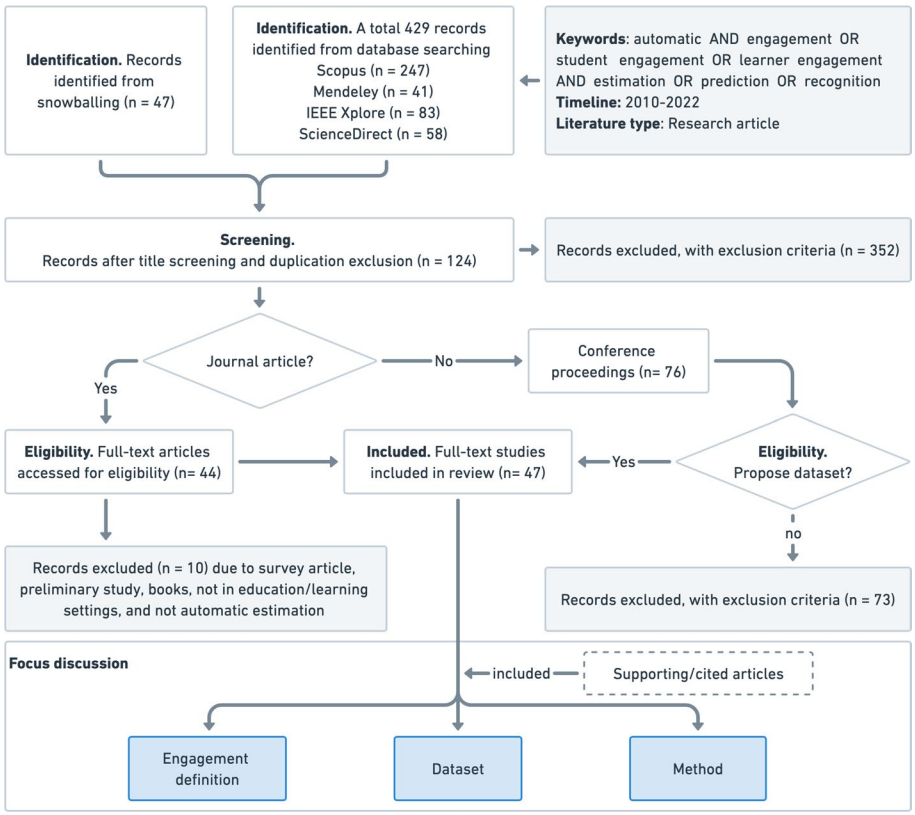

Fig. 1 Illustration of the article selection process adopted from PRISMA flow diagram

addressing the abstract, introduction, methods, results, and discussion. However, for readability reason, we include the discussion in results section, which also is presented to address each RQ, and in conclusions section.

A literature search was carried out based on PRISMA flow diagram with modification by adding eligibility phase. Therefore, there are 4 phases in the flow, i.e., identification, screening, eligibility, and inclusion. We also modify the flowchart by adding initial inclusion criteria (such as keywords, timeline, and literature type), and focus discussion (i.,e., engagement definition, dataset, and method). Figure 1 shows the modified PRISMA flowchart used to select articles in this review.

Identification

The literature search was carried out by selecting research articles from the following electronic databases and libraries: Scopus, Mendeley, IEEE Xplore, and ScienceDirect. The following criteria were used to define the included studies:

- Focused on machine learning-based estimation
- Deployed in education/learning settings
- Journal publications or conference proceedings only if they developed an influential dataset for engagement estimation.

Based on the above criteria, we identified articles that satisfied the following terms: (1) keywords: automatic AND engagement OR student engagement OR learner engagement AND estimation OR prediction OR recognition; (2) publication year: 2010-2022; and (3) literature type: research article, excluding books, magazines, news articles, and posters. Additionally, to obtain more references, we used the snowballing approach by searching Google Scholar. A total of 429 articles were obtained in the identification phase according to the aforementioned search terms.

Screening

In this phase, duplicate articles were excluded. Then, the titles and abstracts were scrutinized to determine whether they met the review criteria. The exclusion criteria included systematic reviews, surveys, and preliminary works (e.g., only report designs). With the exclusion criteria, 352 articles were excluded, yielding 124 articles.

Eligibility

Journal articles and conference proceedings were assessed for eligibility in this phase. The titles, abstracts, main contents, and conclusions were examined to ensure that they met the inclusion criteria. In addition to the exclusion criteria mentioned in the screening phase, we also excluded articles that did not focus on automatic engagement estimation or were not related to education/learning settings. Even though face detection/recognition is a component of engagement estimation in some cases, we excluded articles that focused more on face detection/recognition than on engagement estimation.

A total of 10 journal articles were excluded in this phase according to the exclusion criteria. For the conference proceedings, only articles that proposed an influential dataset for engagement estimation were included. With this condition, 73 out of 76 articles were excluded.

Inclusion

Finally, a total of 47 articles were selected, including 44 journal articles and 3 conference proceedings. In this review, we focused on three main topics: engagement definitions, datasets, and methods. In the discussion section, we also present some supporting articles with citations in the literature.

Results and discussions

Forty-seven articles were selected for this review. The articles were published between 2010 and 2022, although no included articles were published in 2013. However, research on automatic engagement estimation in education/learning settings has increased in recent years (Fig. 2). In particular, in 2021, 14 articles (29.79%) on this topic were published (doubled from the previous year) following the outbreak of the COVID-19 pandemic, which started in 2020.

Among the selected articles, 3 were published in conference proceedings, and the remainder were published in 33 different journals. Most of the journal articles were published in *IEEE Transactions on Affective Computing* ($n = 9$; 19.15%), *Computers & Education* ($n = 3$; 6.38%), and *Applied Intelligence* ($n = 2$; 4.26%), as shown in Fig. 3.

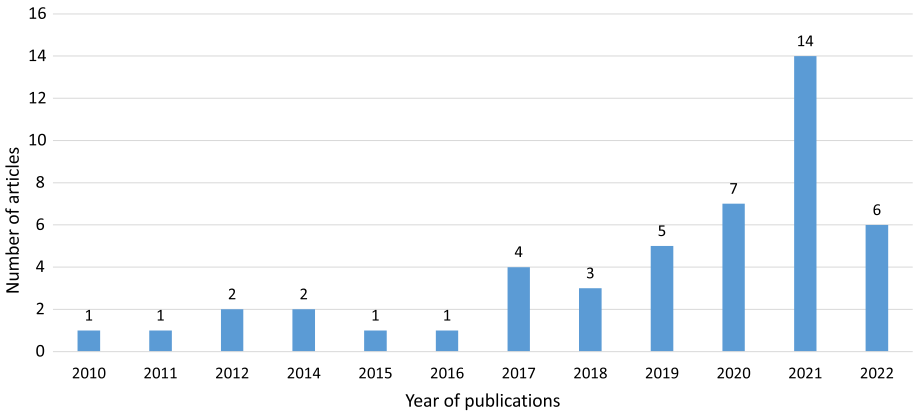

Fig. 2 Number of reviewed articles per year

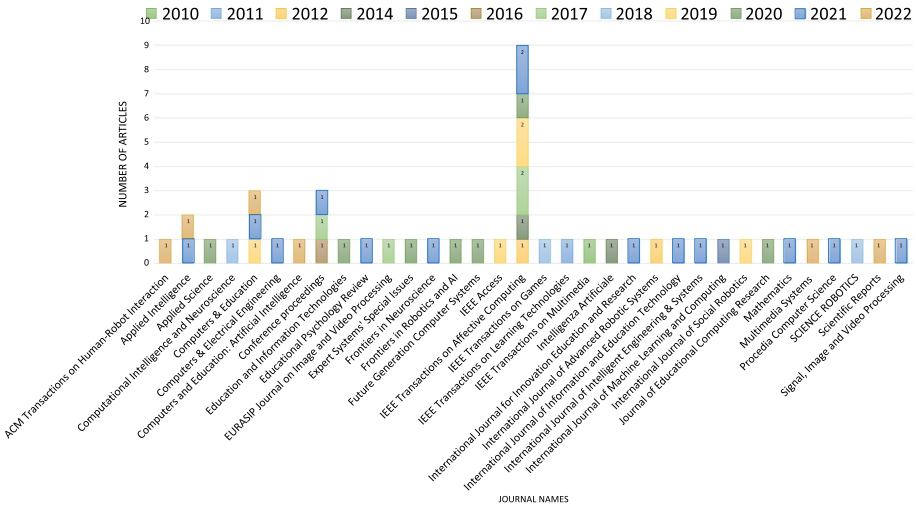

Fig. 3 Number of reviewed articles per journal

Interestingly, research on automatic engagement, which is aimed at education and learning settings, has been conducted in several research domains, including human-human interactions (HHIs), human-robot interactions (HRIs), human-computer interactions (HCIs), and embodied conversational agents (ECAs). From the reviewed articles, we also noted some studies based on data from offline classrooms. Therefore, we added the classroom as a separate research domain in this review.

As shown in Fig. 4, research on automatic engagement estimation in education/learning settings was dominated by HCI ($n = 28$; 59.57%), followed by HRI ($n = 10$; 21.28%) and Classroom ($n = 7$; 14.89%) (see Appendix Table 2).

RQ1: how should the type of engagement to be measured be defined?

In engagement estimation studies, the definition of engagement varies considerably. The definition of engagement depends on the main focus of the study (Christenson et al., 2012; Keen 2009). In educational or learning contexts, we found that the three main research domains depended on the engagement stimuli: HCIs, HRIs, and ECAs, HHIs.

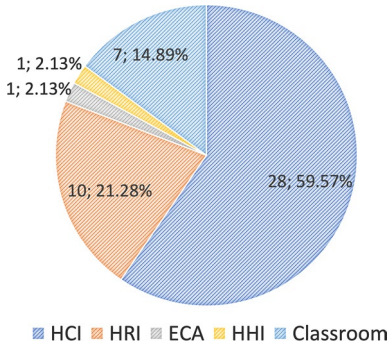

Fig. 4 Number of reviewed articles per research domain

Although HRI research is novel in the field of education, robots can assist humans in different learning processes, such as helping children learn cognitive and social skills and supporting educators teaching difficult concepts (Sharkawy 2021). Robots can not only assist in learning processes but can also measure and increase learner engagement (Celiktutan et al., 2019; Rudovic et al., 2018b; Del Duchetto et al., 2020). HRI researchers defined engagement via two approaches. The first approach defines engagement as a process during interactions that combines verbal and nonverbal communication between two (or more) partners, and the second approach defines engagement as an interaction quality metric.

Moreover, researchers who focused on ECAs (Peters et al., 2005) and intelligent tutor systems (ITSs) (D'Mello et al., 2007) viewed engagement as a value that indicates how likely a person is to remain with their partner and continue an interaction.

Furthermore, engagement estimation research in the field of HCI defined engagement based on engagement cues in computer-based learning, such as learners watching videos, writing, and playing educational games, or in classroom recordings (Whitehill et al., 2014; Monkaresi et al., 2017; Sumer et al., 2021).

This inconsistent definition of engagement in the literature due to the lack of consensus and taxonomy for learning engagement (Yue et al., 2019) may cause confusion for new researchers in this field. To address this challenge, we introduce a taxonomy for engagement and systematically review the definition of engagement used in the selected articles (Fig. 5). As a baseline, we follow the definition of engagement in education and learning environments proposed by Fredricks et al., (2004), which has been widely used in engagement research (Wolters and Taylor 2012; Finn and Zimmer 2012; Greene 2015; Xie et al., 2019; Azevedo 2015).

Engagement is associated with internal states constructed by various cues and may not be visually apparent. Fredricks et al. (2004) divided engagement into three categories: behavioural, emotional, and cognitive engagement. Although, in this definition of engagement, the components to construct each type of engagement overlap considerably, as shown in Fig. 5.

Behavioural engagement describes learners' participation in learning and tasks (Fredricks et al., 2004). In classroom settings, behavioural engagement is shown by actively participating in class, such as by asking questions or displaying attention and concentration (Sumer et al., 2021). *Emotional* engagement refers to learners' affective

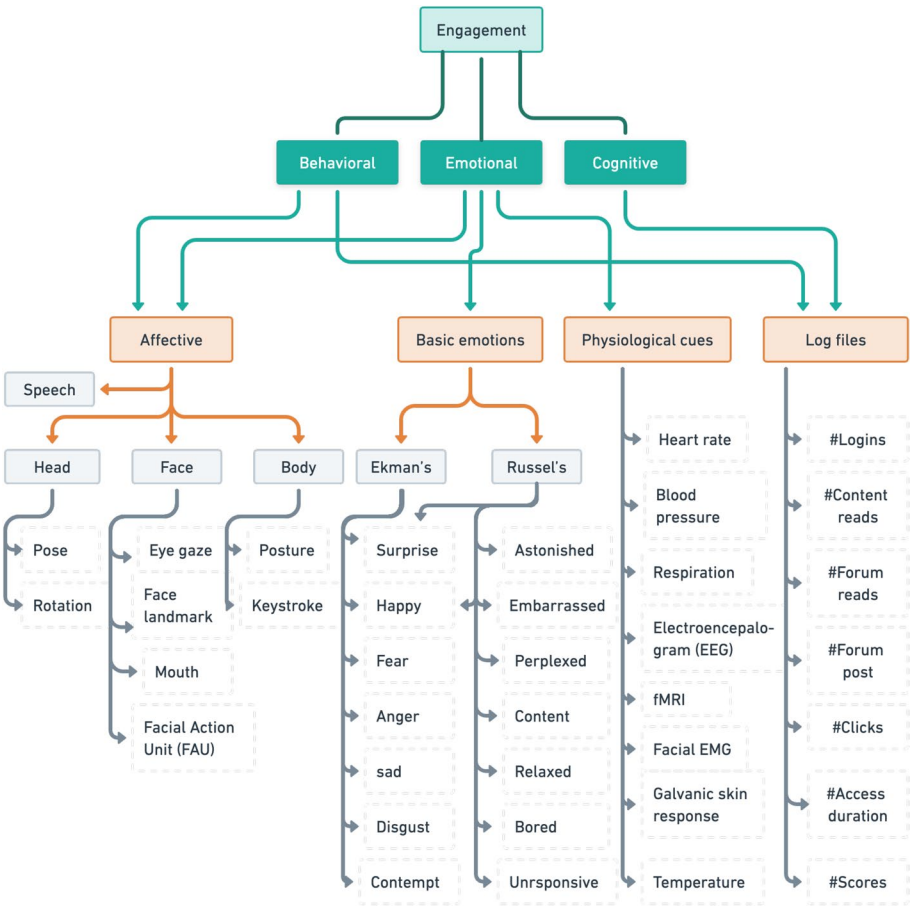

Fig. 5 Taxonomy of engagement definition and its components

reactions in the classroom or during learning, including interest, boredom, happiness, sadness, and anxiety (Fredricks et al., 2004). *Cognitive* engagement, which is also referred to as self-regulation, incorporates learners' psychological investment in learning, including their flexibility in problem solving, learning motivation, and coping mechanisms when faced with failure.

The components for assessing engagement include effort, attention, and persistence for *behavioural* engagement; various emotional reactions (such as anger, surprise, disgust, enjoyment, fear, and sadness (Ekman and Friesen 1978)) to the learning materials for *emotional* engagement; and metacognitive strategies, namely, how learners set goals, plan, and organize their study efforts, for *cognitive* engagement (Fredricks et al., 2004).

In developing automatic engagement estimation methods, these components can be obtained with several modalities (Table 2), such as *log files*, which include information related to learner performance, reaction times, and errors (Cerezo et al., 2016; Okubo et al., 2017; You 2016); *affective cues*, including face and body analyses from video/images (Whitehill et al., 2014; Bosch et al., 2016; Bosch 2016); and *physiological cues*, such as galvanic skin responses (Di Lascio et al., 2018; McNeal et al., 2020), electroencephalograms (EEGs) (Poulsen et al., 2017; Bevilacqua et al., 2019),

heart rates (Darnell and Krieg 2019; Monkaresi et al., 2017), and combinations of these cues (D'Mello et al., 2017).

The engagement level can be determined by grouping emotions according to Ekman's basic emotions (Ekman and Friesen 1978) or Russel's model (Russell 1980). For example, Altuwairqi et al., (2021a) suggested that 'surprised' indicates *strong* engagement; 'enthusiastic', 'excited', and 'nervous' indicate *high* engagement; 'satisfied' and 'happy' indicate *medium* engagement; and 'bored' indicates *low* engagement. Other behaviours, such as not looking at the computer and playing with hair, are classified as disengagement. For two-level classification, strong, high, and medium engagement are grouped into the high engagement class, while low and disengagement are grouped into the disengagement class. In addition, Olivetti et al., (2019) divided engagement level into three classes based on the first and fourth quadrants of Russell's model: **Class 1** included bored, relaxed, and unresponsive; **Class 2** included happy, attentive, content, and perplexed; and **Class 3** included surprised, astonished, and embarrassed.

Consulting the taxonomy, we then reviewed the definition of engagement with a two-step approach. First, we examined the modalities used in each article and how the engagement level was determined. The articles included three common engagement modalities: affective cues (including audio and visual), physiological cues, and log files that were annotated to determine engagement. Some works used publicly available datasets or facial expression tools that already included basic emotion labels. Therefore, we included basic emotions in the taxonomy at the same level as the other modalities to further define the type of engagement (i.e., behavioural, emotional, or cognitive). Note that one engagement cue does not exclusively correspond to one engagement type, as previously discussed.

For example, Apicella et al. (2022) estimated emotional and cognitive engagement with a physiological sensor, i.e., EEG signal acquisition, because the type of stimuli considered during data collection was related to internal emotions and the cognitive task. In this case, two types of stimuli, namely, social feedback and background music, which were organized based on Russel's four quadrants, were used to estimate emotional engagement, while a cognitive task (Continuous Performance Test) was used to estimate cognitive engagement.

Moreover, Goldberg et al. (2021) analysed three types of engagement with one modality, namely, videos recorded in an offline classroom. The behaviour of the students (on- or off-task) in the videos and a knowledge test presented during the lecture were used to estimate the behavioural and cognitive engagement levels, while facial features were extracted from the video to analyse emotional engagement. Therefore, in addition to the engagement cues used, defining what type of engagement is being measured depends on what stimuli were presented to the participant during data collection and what physical or cognitive behaviours were observed.

Overall, most of the selected articles analysed emotional engagement ($n = 40$; 65.57%) with affective cues ($n = 38$; 57.58%), including visual (from videos, which show facial, body, and head information) and audio (speech) cues (Figs. 6 and 7) (See Appendix Table 2).

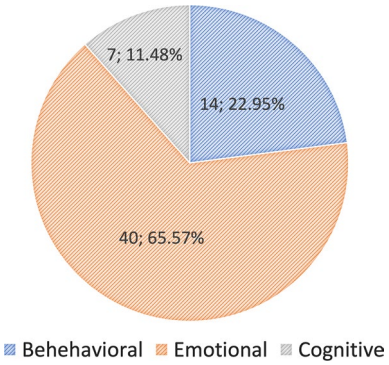

Fig. 6 Pie chart of the engagement types estimated in the selected articles

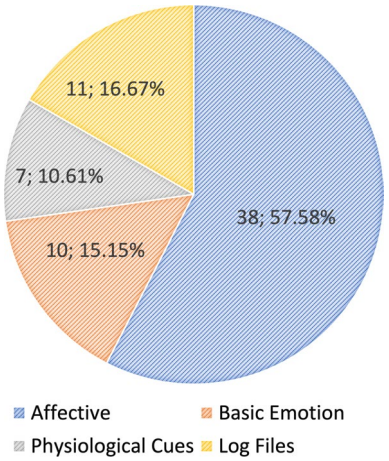

Fig. 7 Pie chart of the engagement cues measured in the selected articles

RQ2: what datasets are suitable for developing automatic engagement estimation methods?

Adequately labelled data and a sufficient amount of data that includes as many generalized variables as possible are important criteria for developing automatic engagement estimation methods. Automatic engagement estimation approaches can be developed by using publicly available datasets or self-collected datasets. Publicly available datasets are open, freely downloadable and may have some terms and conditions, such as use in only research contexts or with consent from the authors. Moreover, self-collected datasets (also referred to as non-public datasets) are built according to specific tasks and cannot be publicly shared due to privacy policies and ethics.

In contrast to emotion recognition datasets, which are typically labelled based on Ekman's basic expressions (e.g., anger, disgust, fear, happiness, sadness, surprise, and neutral), there are only a few publicly available engagement datasets, i.e., datasets that include 'engagement' in their labelling process. However, as shown in the taxonomy of engagement estimation (Fig. 5), an emotion recognition dataset can be used for automatic engagement estimation by modifying labels or by introducing other measurement metrics to define engagement types. In this article, we refer to datasets that are used in the automatic engagement estimation literature even though they have no

straightforward engagement labels as *engagement-related datasets* and datasets that have 'engagement' label as *engagement datasets*.

The selected articles include four engagement-related datasets and three engagement datasets that are publicly available. The public engagement-related datasets include: (1) the NVIE dataset[1] (Wang et al., 2010), (2) BAUM-1 (Zhalehpour et al., 2017), (3) the MASR dataset, which is used in Psaltis et al. (2016) but was proposed in Psaltis et al. (2016), and (4) AffectNet (Mollahosseini et al., 2019). The public engagement datasets include: (1) DAiSEE[2] (Gupta et al., 2016), (2) UE-HRI[3] (Ben-Youssef et al., 2017), and (3) MHHRI[4] (Celiktutan et al., 2019) (see Appendix Table 3).

DAiSEE is one of the most popular publicly available engagement datasets used in the literature (Pabba and Kumar, 2022; Liao et al., 2021; Ma et al., 2021; Thiruthuvanathan et al., 2021; Mehta et al., 2022). Another popular publicly available engagement dataset is the Emotion Recognition in the Wild (EmotiW) dataset. This dataset was excluded from this review because the dataset is being continuously updated; however, EmotiW 2018 (Dhall et al., 2018) and 2020 (Dhall et al., 2020), are accessible for academic research (ACM International Conference on Multimodal Interaction 2020, 2020).

The data in DAiSEE and EmotiW were collected in 'in-the-wild' environments, where participants contributed to the data collection process by recording themselves showing their upper body while watching learning videos. The participants could join from anywhere, and no camera or lighting specifications were considered. Therefore, the quality (e.g., illumination, background noise, and occlusion) of the videos varies. Although in-the-wild data have considerable variations, they are believed to be the closest to real-world conditions (Gupta et al., 2016; Dhall et al., 2018, 2020).

Despite the ease and amount of available data, DAiSEE, EmotiW, and other publicly available datasets were collected with participants of certain ethnicities, which may not be appropriate for all target subjects. Moreover, 'in-the-wild' data may be difficult to process due to the large variations. Therefore, most engagement studies build custom engagement datasets that address the requirements of their model or system (Appendix Table 3). However, because data collection is costly and time-consuming, the amount of data collected may be insufficient. In such cases, self-collected data can be combined with engagement-related datasets or transfer learning data to enhance the estimation performance.

Transfer learning is a type of fine-tuning, which is briefly described in Sect. "Fine-Tuning and Transfer Learning Techniques". In general, transfer learning involves using a pre-trained neural network on a large dataset to extract features to use on tasks with smaller datasets. Some image datasets used for transfer learning include FER-2013 (Goodfellow et al., 2013), VGGFace (Parkhi et al., 2015), VGGFace2 (Cao et al., 2018), FaceNet (Schroff et al., 2015), AffectNet (Mollahosseini et al., 2019), 300W-LP and AFLW2000 (Zhu et al., 2016), JAFFE (Lyons et al., 2002), CK+ (Lucey et al., 2011), and RAF-DB (Li et al., 2017) (see Appendix Table 3).

[1] A natural visible and infrared facial expression database for expression recognition and emotion inference.

[2] User engagement in spontaneous human-robot interactions.

[3] Dataset for affective states in E-environment.

[4] Multimodal human-human-robot interactions dataset for studying personality and engagement.

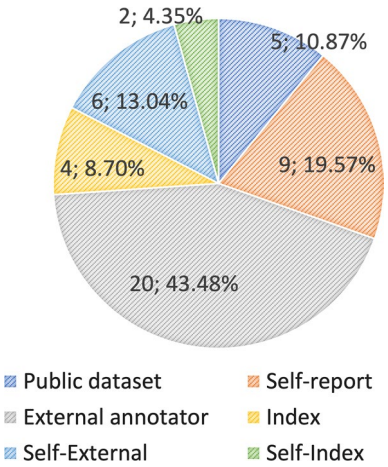

Fig. 8 Pie chart of the engagement measurements and annotations used in the selected articles

Engagement measurement

There are various approaches for measuring engagement, including self-reports, experience sampling techniques, teacher ratings, interviews, and observations (Fredricks and McColskey 2012). In addition, different indices (such as performance indices, number of clicks, and sensor data) have been used to assess engagement (Yue et al., 2019; Apicella et al., 2022; Yun et al., 2012). However, external observations, self-reported measures and ratings are commonly used to measure engagement (Whitehill et al., 2014; Christenson et al., 2012). Moreover, most publicly available engagement datasets were collected based on external observations by external annotators ($n = 20$; 43.48%) (Fig. 8).

Self-reported measures are cheaper and easier to collect than external observations, which require more personnel to measure engagement (Christenson et al., 2012). Self-reports can be performed by self-annotating or completing questionnaires related to self-engagement (O'Brien and Toms 2010). However, self-reported measures are prone to Dunning-Kruger effects, as people are biased in recognizing self-competence (Kruger and Dunning 1999; Pennycook et al., 2017). In addition, these measures dependent strongly on participant compliance and diligence (Eisele et al., 2022). The bias associated with self-reported measure was also observed by Ramanarayanan et al. (2017b; a).

Furthermore, observational measures limit the judgement quality of learners' actual effort, participation, or thinking (Fredricks et al., 2004; Peterson et al., 1984). An external observer is an overhearer (Schober and Clark 1989) that may not consider nonverbal behaviours as signs of engagement. For example, learners who are judged to be on-task or engaged by observers may not actually be thinking about the learning material. In contrast, some learners who appear to be off-task or unengaged may be attempting to understand or relate new ideas to what they have learned (Peterson et al., 1984). In addition, in terms of cognitive engagement, cognition is not easily observable and must be inferred from behaviours or assessed according to performance or self-reported measures (Fredricks et al., 2004; Winne and Perry 2000).

Alternatively, index measurements and combination approaches have been applied to reduce bias. Among the selected articles, four (8.70%) studies used index measurements,

six (13.04%) studies combined self-reported measures with external observations, and two (4.35%) studies combined self-reported measures with some index.

Trindade et al. (2021) performed calculations on log data from courses in Moodle to evaluate engagement. Similarly, Hasnine et al. (2021) calculated concentration indices, Apicella et al. (2022) combined self-reported measures with performance indices, and Yue et al. (2019) combined self-reported measures with quiz scores to assess engagement.

Annotations

Annotation is a crucial step in building a good dataset. Single data points can be annotated manually by one or multiple annotators or by using a framework (Chi and Wylie 2014) or annotation tools such as CARMA (Girard 2014), ANVIL (Kipp 2008), NOVA (Baur et al., 2015), and ELAN (Wittenburg et al., 2006; Brugman and Russel 2004), as shown in Table 3.

To determine whether the labels are consistent, an agreed-upon final label must be determined by several annotators, for example, by using Cohen's kappa value (Wang et al., 2010; AlZoubi et al., 2012; Whitehill et al., 2014; Zhalehpour et al., 2017; Ashwin and Guddeti 2020a, b). Cohen's kappa has also been used to evaluate the efficiency of classifiers for multiclass and imbalanced data (Thiruthuvanathan et al., 2021).

The final label can also be determined by measuring intraclass correlations (ICCs) (Goldberg et al., 2021; Rudovic et al., 2018b) or by applying the majority-vote aggregation technique (Yun et al., 2020; Pabba and Kumar 2022; Zhalehpour et al., 2017). Highly consistent labelled data usually indicate a high degree of credibility (Zhang et al., 2020).

Labelling issues

Visual computer vision-based engagement estimation datasets encounter several challenges, such as various camera angles and image quality (illumination, background, occlusion, etc.). In addition, the difficulty in capturing subtle changes in visual appearance leads to mislabelling issues. For example, one video clip may show more than one engagement state annotated as one state. As a result, some frames may be mislabelled, potentially influencing the frame-by-frame estimation process (Yun et al., 2020). Frame-based labelling is viewed as the easiest solution. However, this approach lacks continuous labels, which provide more precise information (Sumer et al., 2021). To address this issue, temporal dynamics features need to be extracted (Yun et al., 2020).

However, in some cases, some frames are more significant for determining engagement levels, while other frames can mislead the final estimation result (Zhu et al., 2020). One solution for addressing this problem is applying an attention mechanism. The attention mechanism in deep learning directs attention to effectively choose important frames (Vaswani et al., 2017; Winata et al., 2018).

Another labelling issue is false interpretation. For example, learners may be engaged regardless of where they are looking, and observers might label a learner who looks down as disengaged while the learner is actually thinking or processing the learning material. Especially in higher grade levels, learners may show/hide their engagement, and engagement cues may thus be more difficult to identify (Lufi and Haimov 2019). Moreover, age can affect attention levels (Lufi and Haimov 2019). Therefore, collecting

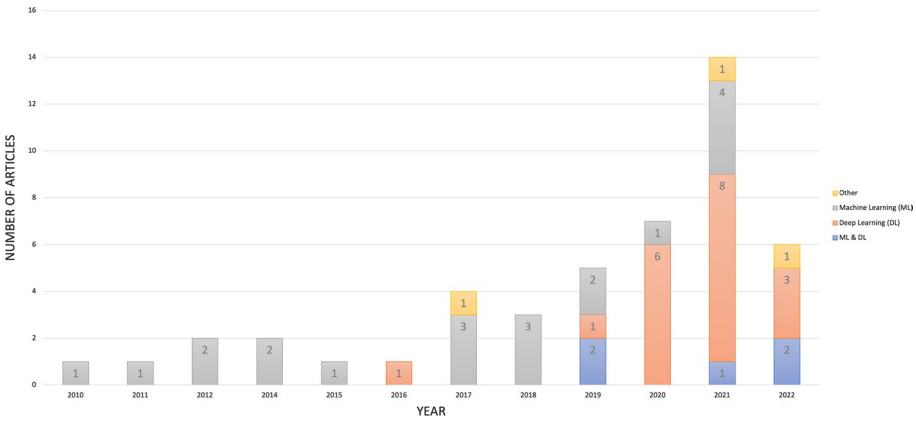

Fig. 9 The general method used in the selected articles

an engagement dataset that represents learners' authentic internal states is a challenging task.

RQ3: what automatic engagement estimation methods have been developed in the literature?

Machine learning, which is a subset of artificial intelligence (AI), is known for its capability to acquire knowledge to make decisions by extracting patterns from raw data (Goodfellow et al., 2016). Machine learning techniques have been applied in various fields, including agriculture, transportation, business, and education. Machine learning has led to the development of affective computing methods that automatically recognize human emotions and behaviours (Schuller 2015; Kratzwald et al., 2018; Zhao et al., 2019; Rouast et al., 2021), supporting the advancement of artificial intelligence in education applications (Chen et al., 2020; Ouyang and Jiao 2021). Therefore, in general, automatic engagement estimation methods are referred to as machine learning (ML)-based algorithms.

Since machine and deep learning methods are the most commonly used approaches for developing automatic engagement estimation tools in the literature (Fig. 9), in this section, we briefly discuss the pre-processing steps and estimation methods (classification or regression). We classified the estimation methods as classic machine learning and deep learning techniques.

Deep learning is a subset of machine learning. Both techniques work by mapping raw data features to extract the desired information. Nevertheless, it may be difficult for computers to extract features from raw data with large variations, and these features may be identifiable only using a nearly human-level understanding of data (Goodfellow et al., 2016). Therefore, classic machine learning methods require hand-designed features. Moreover, deep learning approaches reduce the desired complicated mapping into a series of nested mappings that can be described by layers (Goodfellow et al., 2016). For example, to identify image features, the input is presented as a visible layer. Then, the next layers, namely, the hidden layers, divide the image into smaller maps such as edges, corners and contours, object parts, and finally, the object identity. Figure 10 depicts a Venn diagram showing how deep learning is distinguished from classic machine learning.

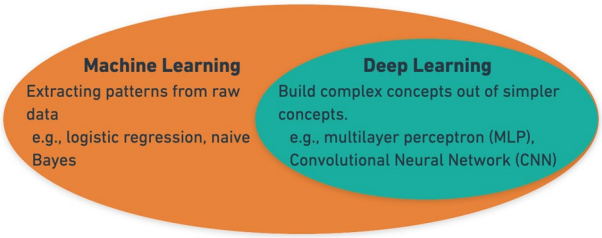

Fig. 10 The difference between machine learning and deep learning

Table 1 Face recognition tools for face detection and feature extraction

Tools name	Used in
OpenFace (Baltrusaitis et al. 2016, 2018)	(Kaur et al. 2018; Rudovic et al. 2018b; Goldberg et al. 2021; Ma et al. 2021; Wu et al. 2020; Zhang et al. 2019; Zhu et al. 2020; Thong Huynh et al. 2019; Li et al. 2021; Engwall et al. 2022)
OpenCV	(Yang et al. 2018; Wang et al. 2020; De Carolis et al. 2019; Bhardwaj et al. 2021; Hasnine et al. 2021)
Dlib	(Hasnine et al. 2021; Mehta et al. 2022)
OpenPose (Cao et al. 2017)	(Vanneste et al. 2021; Zheng et al. 2021; Wu et al. 2020; Zhu et al. 2020)
RetinaFace (Deng et al. 2019)	(Sumer et al. 2021)
FasterRCNN (Ren et al. 2015)	(Rudovic et al. 2018a)
faceAPI	(Castellano et al. 2012)
Affectiva API in iMotion	(Dubovi 2022)

Pre-processing

Before data can be fed into a network, the raw data must be pre-processed to extract the features. Video/image-based data can be pre-processed with face detection, tracking, and cropping techniques (Yun et al., 2020). Alternatively, statistical values can be extracted to obtain representation information from features in a given time window (Hernandez et al., 2013; Sanghvi et al., 2011; Yun et al., 2020). Statistical rules such as sum, max, min, and mode can be utilized to aggregate meaningful information as input for classifiers, including support vector machines (SVMs) and neural networks (Yun et al., 2020).

Face detection and feature extraction Appearance-based features can be divided into two categories: low-level features and high-level features. Low-level features include the information generated in each video frame in a given time window. In particular, HCI engagement research has adopted low-dimensional geometry and appearance descriptors as features (Sumer et al., 2021; Whitehill et al., 2014). Additional low-level features include local binary patterns in three orthogonal planes (LBP-TOP), Gabor features, and box filters (BFs) (Li and Deng 2020).

High-level features are features extracted by aggregating low-level features (Yun et al., 2020), such as facial action units (FAUs) and head poses. Facial features and head poses are some of the most commonly used features for determining engagement and attention (Akker et al., 2009; Ba and Odobez 2006; Dong et al., 2010; Voit and Stiefelhagen 2008; Zhang et al., 2007). These features can be extracted statistically or by using facial recognition tools, as shown in Table 1.

OpenFace is a popular computer vision toolkit for extracting facial features, including in automatic engagement estimation research (Table 1). OpenFace implements multitask cascaded convolutional networks (MTCNNs) (Zhang et al. 2016) for face detection, constrained local models (Baltrusaitis et al., 2013; Zadeh et al., 2017) for landmark detection and tracking, eye rendering (Wood et al., 2015) for eye gaze estimation, and cross-dataset learning and person-specific normalisation for facial action unit (FAU) detection. In addition, the OpenCV[5] face detection library (Haar Cascade (Viola and Jones 2004, 2001; Schmidt and Kasiński 2007)) and Dlib library for face and landmark detection are widely used. The mean shift-based object tracker in OpenCV can also be used for face tracking. Furthermore, in HRI, face recognition can be performed by utilizing the software development kit (SDK) built into the robot, for example, the NAOqi People Perception in the Pepper robot (Ben-Youssef et al., 2021). Interested readers are referred to (Wang and Deng 2021) for an in-depth explanation, especially deep learning-based face recognition.

Data augmentation Data augmentation is the process of creating new data based on real data without changing the original data. For image inputs, data augmentation can be performed by flipping (horizontally or vertically), cropping, scaling, or translating/rotating the images. As a result, the sampling rate for the input can be increased by adding the augmented data to the original dataset (Shen et al., 2022; Ashwin and Guddeti 2020b; Pabba and Kumar 2022).

Feature selection Feature selection not only determines the optimal set of features but also ranks and compares the most discriminative features. Some feature selection methods include F-scores (Chen and Lin 2006), RELIEF-F (Whitehill et al., 2014), DeepLift (Rudovic et al., 2018b), and recursive feature elimination random forests (RFE-RFs). Alternatively, ANOVA can be used to analyse the significance of labelled features (Schiavo et al., 2014).

Dimensional reduction Dimensional reduction is the process of decreasing the dimension of the input feature to prevent overfitting (Yun et al., 2020). Dimensional reduction can be applied to a dataset before the data are fed into the network. Some dimensional reduction methods include principal component analysis (PCA) (Sumer et al., 2021; Wang et al., 2010) and forward feature selection (FFS) (AlZoubi et al., 2012). However, dimensional reduction can also be performed by layer reduction using various pooling layers (max, average, and variance pooling, 1x1 convolutional layers) when a convolutional neural network is utilized (Yun et al., 2020).

Addressing imbalanced data One major issue with engagement datasets is imbalanced data that are severely skewed towards the majority class (Yun et al., 2020). Imbalanced class labels often occur because disengagement is rarely observed in continuous labelling. Many methods have been proposed to address this issue (Galar et al., 2012; Chawla et al., 2002; García et al., 2012; Dresvyanskiy et al., 2021). There are three categories of resampling techniques (Ben-Youssef et al., 2021): (1) undersampling methods, which aim to balance class distributions by eliminating majority class examples; (2) oversampling methods, which generate minority class examples, e.g., the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002); and (3) hybrid methods that combine

[5] https://opencv.org/.

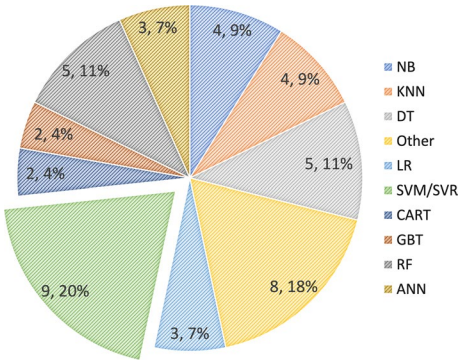

Fig. 11 Pie chart of the use of classic machine learning methods for automatic engagement estimation

both sampling methods (García et al., 2012; Chawla et al., 2002). Moreover, continuous scales may be discretized into groups (Rudovic et al., 2019a, b), and weighting techniques (Dresvyanskiy et al., 2021; Lin et al., 2017) have also been used to address this problem.

Classic machine learning methods

Engagement is estimated by calculating probabilities. To calculate the engagement probability, several classic machine learning methods can be utilized, such as the support vector machine (SVM) and its variations (including support vector regression (SVR)), naive Bayes (NB), decision trees (DTs), logistic regression (LR), clustering techniques (e.g., K-nearest neighbour (KNN)), and random forest (RF). These machine learning techniques are conveniently available in machine learning toolboxes such as Waikato Environment for Knowledge Analysis (WEKA) (Witten and Frank 2005) (as used in (Cocea & Weibelzahl 2011; Monkaresi et al., 2017; Ribeiro Trindade and James Ferreira 2021)), the Computer Expression Recognition Toolbox (CERT) (Littlewort et al., 2011) (as used in (Whitehill et al., 2014)), and the MATLAB library ((Chatterjee et al., 2021; AlZoubi et al., 2012).

Between 2010 and 2022, classic machine learning methods dominated the automatic engagement estimation literature (Fig. 9), especially SVMs (Fig. 11). Note that some of the selected articles examined more than one algorithm. Therefore, the totals in Fig.11 do not correspond to the number of selected articles (see Appendix Table 4).

Deep learning methods

With the development of deep learning, research on automatic engagement estimation has applied these techniques to improve the estimation performance (Fig. 9). In this section, we briefly introduce some deep learning methods, including those used in the selected articles. For a more detailed explanation on deep learning techniques (especially for face recognition), interested readers are referred to (Wang & Deng, 2021; Fuad et al., 2021; Li & Deng, 2020).

Multilayer perceptron (MLP) The multilayer perceptron (MLP), also called the feedforward neural network or deep forward network, was one of the first deep learning algorithms. The MLP is a mathematical function that is formed by combining many simpler functions to map some set of input values to output values (Goodfellow et al., 2016). An MLP consists of at least three layers of nodes, i.e., the input $f^{(1)}$, hidden $f^{(2)}$, and output $f^{(3)}$ layers, to define the mapping $y \approx f^*(\mathbf{x}) = f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$. The first and last

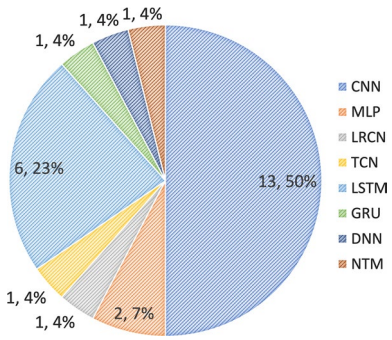

Fig. 12 Pie chart of the use of deep learning methods for automatic engagement estimation

layers are called the input and output layers, respectively, while the number of hidden layers may be varied, which determines the *depth* of the model. Furthermore, each layer may contain more than one unit depending on the number of inputs and outputs. This algorithm has also been used in automatic engagement estimation for performance comparison with other algorithms (Ben-Youssef et al., 2019; Sumer et al., 2021; Rudovic et al., 2018b).

Convolutional neural network (CNN) A convolutional neural network (CNN) is a specialized kind of deep learning (DL) algorithm for processing data that employs mathematical linear operations known as *convolutions* as opposed to matrix multiplication (Goodfellow et al., 2016). The convolution operation is typically denoted with an asterisk: $x'(t) = (x * w)(t)$, where x' is the *feature map*, i.e., the estimated value from the convolution of the *input* x with a *kernel* w at time t (Goodfellow et al., 2016).

CNNs are currently one of the most popular methods in different fields (Fig. 12). This technique has been widely used in various computer vision applications, including image classification (He et al., 2016), semantic segmentation (Noh et al., 2015), object detection (Szegedy et al., 2015), face recognition (Parkhi et al., 2015), spatiotemporal feature learning (Tran et al., 2015; Husain et al., 2016; Ji et al., 2013; Yun et al., 2020; Rudovic et al., 2018a; Abedi and Khan 2021; Ashwin and Guddeti 2020c; Yue et al., 2019), and automatic engagement estimation (see Appendix Table 4).

CNNs are popular because they can be highly modified and pretrained. Some CNNs include AlexNet (Krizhevsky et al., 2017), i3D (Carreira and Zisserman 2017), VGG16 (Simonyan and Zisserman, 2014), and ResNet (He et al., 2016; Szegedy et al., 2015).

The inputs to a CNN are usually greyscale or RGB images. The use of multiple small filtering kernels allows the network to extract more discriminative features because multiple small kernels are easier to optimize than one large filter kernel (Mohamad et al., 2020; Wang et al., 2020). However, CNNs have some crucial issues, such as large training times, gradient vanishing due to the use of deep networks, and a large number of parameters (Thiruthuvanathan et al., 2021).

Recurrent neural network (RNN) A facial expression changes through three stages, i.e., onset, apex, and offset (Liu et al., 2014). In recurrent neural network (RNN) algorithms for engagement estimation, time-series images are more reasonable than static images as input since time-series present sequence-related task information (Jordan 1990). RNNs capture information at earlier and later time steps by remembering each

piece of information over time (Sharkawy 2020). Therefore, this algorithm has become a more popular automatic engagement estimation method (see Appendix Table 4).

Some types of RNNs include long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) (Yue et al., 2019; Ben-Youssef et al., 2019; Del Duchetto et al., 2020; Liao et al., 2021; Sumer et al., 2021; Engwall et al., 2022), gated recurrent units (GRUs) (Ben-Youssef et al., 2019), and network Turing machines (NTMs) (Qiao and Bi 2020; Ma et al., 2021).

However, despite advantages such as considerable computational power in temporal processing models and applications, in practice, RNNs are difficult to train due to network instability (Sharkawy 2020). Moreover, the networks may suffer from short-term memory issues if the input sequences are too long. Thus, RNNs may have difficulty capturing earlier time step information due to vanishing gradients (Sharkawy 2020).

Therefore, the attention mechanism was introduced to learn to associate the elements in sequence C with the elements in the output sequence (Bahdanau et al., 2014). The attention mechanism essentially determines a weighted average that is used to focus on specific parts of the input sequence at each time step (Goodfellow et al., 2016). Although the attention mechanism was originally introduced in the context of machine translation (Bahdanau et al., 2014), it has also been utilized in DL applications for automatic engagement estimation (Liao et al., 2021; Sumer et al., 2021; Mehta et al., 2022; Shen et al., 2022).

Other classifiers Other neural network techniques that have been used for automatic engagement estimation include the fuzzy min-max neural network (FMMNN) classifier (Simpson 1992; Gabrys and Bargiela 2000), which was implemented by (Yun et al., 2012) for automatic engagement estimation, the deep belief network (Hinton et al., 2006), which was used in (Dewan et al., 2018), and linear discriminant analysis (LDA) (Apicella et al., 2022; Wang et al., 2010).

Fine-tuning and transfer learning techniques

One fine-tuning technique for addressing insufficient training data is applying transfer learning, which utilizes networks pretrained on a large number of images (Bengio 2011; Wang and Deng 2021). Various models have been trained on large face image datasets. For example, Sumer et al. (2021) used AffectNet (Mollahosseini et al., 2019) and 300W-LP (Zhu et al., 2016), which were trained on ResNet50, for transfer learning. The pretrained models help the engagement estimation network learn general features related to face identification (Yun et al., 2020). As mentioned in Sect. "RQ2", other large datasets that have been used for transfer learning include FER-2013 (Goodfellow et al., 2013), VGGFace (Parkhi et al., 2015), VGGFace2 (Cao et al., 2018), FaceNet (Schroff et al., 2015), AffectNet (Mollahosseini et al., 2019), 300W-LP and AFLW2000 (Zhu et al., 2016).

Performance metrics

To judge the automatic engagement estimation performance, the prediction results should be compared with human judgements in the dataset (Whitehill et al., 2014; Yun

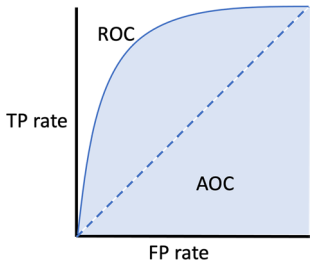

Fig. 13 AUC-ROC curve illustration

et al., 2020). In machine learning pipeline, performance metrics are used to monitor and measure the performance of a model depend on the task. Automatic engagement estimation problem can be seen either as classification or regression task. An engagement estimation is a classification task if the engagement is estimated in discreet class, e.g., low engagement class vs high engagement class. Otherwise, an engagement estimation is a regression task when continuous output desired. Some metrics used to measure the performance of regression task are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which mainly calculating the distance between the predicted and the ground truth.

Classification performance metrics evaluate the estimation model that compares discrete classes, such as accuracy, precision and recall, F1-score, and Area Under the Curve-Receiver Operating Characteristics (AUC-ROC). Moreover, confusion matrix is also used to visualize the ground-truth labels versus the predicted results in a table.

Accuracy metric defines the number of correct predictions (true positive (TP)) divided by the total number of predictions. It is the most common metric for evaluating classification performance due to its simplicity. However, Accuracy may not be reliable when the dataset is severely unbalanced. In a severely skewed dataset, the classifier may not discriminate well despite high accuracy values because the classifier identifies only the most common class.

Alternatively, Precision/Recall (PR) trade-off curve (used in (Leite et al., 2015)) and F1-score (Schiavo et al., 2014) are used to overcome the limitation of Accuracy. Precision determines the performance by calculating the proportion of TP prediction to the total positive prediction (TP + false positive (FP)). Similarly, Recall calculates the TP prediction to the total number of TP and false negative (FN). Meanwhile, F1-score is the harmonic mean between the precision and recall.

Some alternative metrics that are more informative and "imbalance-friendly" include the balanced accuracy, AUC-ROC (Hernandez et al., 2013; Leite et al., 2015) and 2-alternative forced choice (2AFC) (Whitehill et al., 2014).

AUC-ROC visualizes the classification performance based on correct and incorrect classifications (Fig. 13). The ROC curve plotted the trade-off between the TP rate (Recall) to the FP rate. AUC represents the degree or measure of separability between classes as a summary of the ROC curve (Bradley 1997). The AUC scores between $0.7 - 0.8$, $0.8 - 0.9$, and > 0.9 are considered acceptable, excellent, and outstanding, respectively (Mandrekar 2010; Li et al., 2021).

$$Precision = \frac{TP}{(TP + FP)} \qquad Recall\ (TP\ \text{rate}) = \frac{TP}{TP + FN}$$

$$FP\ \text{rate} = 1 - TP\ \text{rate} = \frac{FP}{TN + FP} \qquad F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The 2-alternative forced choice (2AFC) (Mason & Weigel ,2009; Tingfan et al., 2012) is an unbiased estimate of the AUC-ROC curve since it expresses the probability of discriminating true positives (TP) from true negatives (TN). A 2AFC value of 1 indicates perfect discrimination, while a value of 0.5 indicates that the classifier performs at chance levels.

Furthermore, other metrics such as Matthews correlation coefficient (MCC) (Tingfan et al., 2012) and specificity and sensitivity (Yun et al., 2020) are also used in the engagement estimation literature. (see Appendix Table 4).

Conclusion

This article reviewed recent research on automatic engagement estimation in education/learning settings, focusing on work published between 2010 and 2022. In particular, this review examined engagement definitions, datasets, and machine learning-based methods from forty-seven selected articles. The article selection and review methodology were adopted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) model (Page et al., 2021) to answer three research questions:

- RQ1: how should the type of engagement to be measured be defined?
- RQ2: what datasets are suitable for developing automatic engagement estimation methods?
- RQ3: what automatic engagement estimation methods have been developed in the literature?

The results and discussion with the presented information, figures, and tables aim at providing new researchers with insight on automatic engagement estimation to enhance smart learning with automatic engagement recognition methods.

To answer the RQ1, we examined the definitions of engagement used in the selected articles and introduced an engagement definition taxonomy (Fig. 5) as a guide for educators and engagement estimation research, particularly for education/learning purposes. The taxonomy defined three types of engagement: behavioural engagement, emotional engagement, and cognitive engagement. Each engagement type was connected with some engagement cues, including affective cues, physiological cues, log files, and basic emotions. The modalities for obtaining engagement cues were also discussed, including speech cues, visual cues (face, head, and eye gaze), physiological sensor data, and log data.

From the discussion, we found that to define what type of engagement is being measured depends on engagement cues used, what stimulies presented to the participant during data collection, and what physical or cognitive behaviours observed. We believe

that the proposed taxonomy will allow for enhanced research on automatic engagement estimation.

The datasets used in the literature were summarized in this review to address the RQ2. The datasets include publicly available datasets and self-collected datasets. In this review, publicly available datasets were divided into two categories, namely, engagement datasets and engagement-related datasets, to distinguish the availability of engagement labels. The engagement measurement methods and annotations were highlighted because incorrect interpretations in this step leads to severe bias. The number of participants, type of samples, number of annotators, and label information were summarized in a table to provide a reference for building engagement datasets.

Finally, in addressing the RQ3, we discuss machine learning-based methods have been applied to develop automatic engagement estimation approaches in the literature. We found that between 2010 and 2022, classic machine learning algorithms (including support vector machines (SVMs) and decision trees (DTs)) were used more in previous work. However, since 2019, the trend has moved to deep learning algorithms, especially convolutional neural network (CNN)- and recurrent neural network (RNN)-based algorithms.

Limitations and remaining challenges

There is bias in the subjective determination of whether an article was aimed at education/learning settings. For example, some articles appear to be aimed at other purposes, such as therapy for children with autism Rudovic et al. (2018b) or human-robot interactions Ben-Youssef et al. (2019). However, the articles were included if the authors perceived that there was subtle information about a learning activity or the possibility that the proposed action could applied in the education process.

Moreover, the combination of a clear engagement definition, and suitable machine learning methods allows learners' engagement during learning activities to be measured automatically, including human-human interactions, human-computer interactions and human-robot interactions. The estimation performance is especially promising for deep learning-based methods. However, the practicality of the implementation in real education settings is not discussed in this review. Therefore, the implication and application of these automatic engagement estimation methods should be addressed in future work to address various research questions, such as "How does engagement estimation improve learning outcomes?", "What conditions and requirements are needed in automatic engagement estimation applications?", and "In what learning settings can automatic estimation be applied?".

Furthermore, we discuss several remaining challenges, including cognitive engagement, personalized engagement, and machine-learning pitfalls.

Cognitive engagement Table 2 shows that most automatic engagement research has focused on behavioural and emotional engagement and that affective data, especially appearance-based video data, were mostly utilized to estimate engagement. However, cognitive engagement, which can be determined through self-regulated learning or pre-post tests, plays an important role in successful distance learning. Similar to behavioural and emotional engagement, cognitive engagement can be measured using questionnaires

(Li et al., 2021). However, few studies (Table 2) have considered this type of engagement. Therefore, we believe that more engagement cues for cognitive engagement should be developed in future automatic engagement estimation research.

Personalized engagement Various definitions of engagement have been constructed in the field of education. Although engagement can be divided into three types (i.e., behavioural, emotional, and cognitive engagement), conceptualizations of engagement sometimes include only one or two of the three types. All three types can be considered to determine engagement levels (Fredricks et al., 2004). To the best of our knowledge, no research has answered how these engagement types evolve and change over time. Therefore, whether the engagement cues may take different forms depending on the age range, gender, ethnicity, and education level of the participants is unknown.

Moreover, facial physiognomy differences between people with different ethnic backgrounds may result in various distributions of engagement levels (Rudovic et al., 2018a). Several automatic engagement estimations are targeted participants with specific cultures or backgrounds. For example, as shown by Libin and Libin (2004), a child's background, including their cultural or psychological profile, needs to be considered when designing therapeutic strategies.

Network personalization can be achieved using demographic information (culture and gender), followed by individual network layers for each child (Rudovic et al., 2018b). However, it is unknown how engagement estimation results can be generalized in actual applications (Bosch et al., 2016). Thus, the user target must be defined, and the data must be collected from participants with the appropriate cultural background (for example, learners with autism spectrum conditions (ASCs) (Tincani et al., 2009; Conti et al., 2015)) to train the model (Rudovic et al., 2018a). Therefore, automatic engagement estimation, which considers individual differences, remains an open challenge.

Machine learning pitfalls Machine learning (ML) methods have been applied in various fields; however, reproducibility is an issue, as reviewed by Kapoor et al. (2022). The review examined 20 reviews across 17 research fields and found errors in 329 papers that used ML-based methods. While experienced machine learning practitioners are well aware of these errors, researchers in other disciplines may not be (DeepLearning. AI. 2022). Although education research was not included in the review (Kapoor and Narayanan 2022), we found similar issues (such as no training-testing splits, sampling biases, and pre-processing the training and test sets together) in the selected articles (see Appendix Table 4). The misuse of ML can generate invalid results that are irreproducible in implementations in real-world educational settings. Therefore, automatic engagement researchers should be aware of these issues (Kapoor and Narayanan 2022). Furthermore, education experts and ML experts could collaborate on engagement research to develop more effective models (DeepLearning.AI, 2022).

Appendix

Tables 2, 3, and 4.

Table 2 An overview of the selected articles to address RQ1

Author	Journal	Domains				Engagement type			Engagement cues			
		HCI	HRI	A/H	Cls.	B.	E.	C.	Aff.	BE	PC	LF
(USTC-NVIE) Wang et al. (2010)	IEEE Transactions on Multimedia	✓					○		●			
Cocea et al. (2011)	IEEE Transactions on Learning Technologies	✓				○						●
AlZoubi et al. (2012)	IEEE Transactions on Affective Computing			A			○				●	
S-Syun et al. (2012)	International Journal of Advanced Robotic Systems		✓				○		●			
Whitehill et al. (2014)	IEEE Transactions on Affective Computing	✓					○		●			
Schiavo et al. (2014)	Intelligenza Artificiale	✓				○	○		●			
Woo-Han Yun et al. (2015)	International Journal of Machine Learning and Computing	✓					○		●			
(DAiSEE) Gupta et al. (2016)	Conference proceedings	✓					○		●			
Zaletelj et al. (2017)	EURASIP Journal on Image and Video Processing				✓	○			●			
Monkaresi et al. (2017)	IEEE Transactions on Affective Computing	✓					○		●		●	
(UE-HRI) Youssef et al. (2017)	Conference proceedings		✓									
(BAUM-1) Zhalehpour et al. (2017)	IEEE Transactions on Affective Computing	✓										
Hussain et al. (2018)	Computational Intelligence and Neuroscience	✓				○		○				●
Psaltis et al. (2018)	IEEE Transactions on Games	✓					○		●	●		
Rudovic et al. (2018b)	SCIENCE ROBOTICS		✓			○			●		●	
Ninaus et al. (2019)	Computers & Education	✓					○			●		
Yue et al. (2019)	IEEE Access		✓			○	○	○	●	●		●
(AffectNet) Mollahosseini et al. (2019)	IEEE Transactions on Affective Computing	✓					○		●			
Celiktutan et al. (2019)	IEEE Transactions on Affective Computing		✓			○			●		●	
Youssef et al. (2019)	International Journal of Social Robotics		✓				○		●			
Olivetti et al. (2019)	Applied Science	✓					○			●		
Ashwin et al. (2020b)	Education and Information Technologies				✓	○	○		●			
Ashwin et al. (2020a)	Future Generation Computer Systems				✓		○		●	●		
Pabba et al. (2022)	Expert Systems (Special Issues)				✓		○		●	●		
Duchetto et al. (2020)	Frontiers in Robotics and AI		✓				○		●			
Yun et al. (2020)	IEEE Transactions on Affective Computing		✓			○	○		●			

Table 2 (continued)

Author	Journal	Domains				Engagement type			Engagement cues			
		HCI	HRI	A/H	Cls.	B.	E.	C.	Aff.	BE	PC	LF
Zhang et al. (2020)	Journal of Educational Computing Research	✓				○	○		●			●
Liao et al. (2021)	Applied Intelligence	✓					○		●			
Li et al. (2021)	Computers & Education	✓					○	○	●			●
Bhardwaj et al. (2021)	Computers & Electrical Engineering	✓					○		●	●		
Goldberg et al. (2021)	Educational Psychology Review				✓	○	○	○	●			
Chatterjee et al. (2021)	Frontiers in Neuroscience			H			○				●	
Youssef et al. (2021)	IEEE Transactions on Affective Computing		✓			○			●			
Sümer et al. (2021)	IEEE Transactions on Affective Computing				✓		○		●			
Trindade et al. (2021)	International Journal for Innovation Education and Research	✓						○				●
Ma et al. (2021)	International Journal of Information and Education Technology	✓					○		●			
Thiruthvanathan et al. (2021)	International Journal of Intelligent Engineering & Systems	✓					○		●			
Altuwairqi 2021 et al. (2021b)	Signal, Image and Video Processing	✓				○	○		●			●
Vanneste et al. (2021)	Mathematics				✓	○	○		●			
Hasnine et al. (2021)	Procedia Computer Science	✓					○			●		
Delgado et al. (2021)	Conference proceedings	✓					○		●			
Engwall et al. (2022)	ACM Transactions on Human-Robot Interaction		✓				○		●	●		
Mehta et al. (2022)	Applied Intelligence	✓					○		●			
Dubovi et al. (2022)	Computers & Education	✓					○	○	●		●	
Thomas et al. (2022)	Computers and Education: Artificial Intelligence	✓					○		●			●
Shen et al. (2022)	Multimedia Systems	✓					○		●	●		
Apicella et al. (2022)	Scientific Reports	✓					○	○			●	

HCI human-computer interaction; *HRI*human-robot interaction; *A* embodied conversational Agent; *H* human-human interaction; *Cls.* classroom; *B.* behavior; *E.* emotional; *C.* cognitive; *Aff.* affective; *BE* basic emotions; *PC*physiological cues; *LF* log file (including log activity)

Table 3 An overview of the dataset in the selected articles to address RQ2

Dataset	Type	Setting	Stimuli	Participants	Samples	Annotators	Label	Publicity
(USTC-NVIE) Wang et al. (2010)	ER	S &P	3-4 mins emotional and 1-2 mins neutral videos from internet	215 healthy students (157 M and 58 F)	236 apex images. Visible and thermal.	5 EAs & self-report.	**6 basic emotions** (happiness, sadness, surprise, fear, anger, and disgust), average **arousal** and **valence**.	Yes[a]
Cocea et al. (2011)	E	WE	An online course (HTML-tutor) in 7 sessinos	48 users	14 logged events	3 EA	Engaged, disengaged, or neutral	N/A
AlZoubi et al. (2012)	E	S	An intelligent tutoring system with conversational dialogues (AutoTutor) in 45 mins learning session	27 students	20-second interval of biosensor signals	Retrospective self-report	8 affective states: boredom, confusion, curiosity, delight, flow/**engagement**, surprise, and neutral.	N/A
S-Syun et al. (2012)	ER	S	An intellingent robot platform (MERO): greeting, identification, and a questions game.	10 participants	1,400 elements	(humans numerical values)	4 facial expression states: smiling, surprised, neutral and angry	N/A
Whitehill et al. (2014)	E	S	A cognitive skills training software .	34 undergraduate students	10 seconds videos	7EA	Engaged, Not Engaged ([Very engaged,Engaged],[Nominally engaged,Not Engaged])	N/A
Schiavo et al. (2014)	E	S	A video game: single player level, "Operation 40" of "Call of Duty - Black Ops" video game.	22 participants (3 F, 19 M)	12420 samples	self-annotate using ExperienceSampling Method (ESM) (Larson and Csikszentmihalyi 2014)	Neutral, **Engaged**, Stress	N/A
Woo-Han Yun et al. (2015)	E	S	A testing interactive software.	12 Children	2,745 of 30 second video clips	1 EA	4 engagement levels: high/low interest, low/high boredom	N/A
(DAiSEE) Gupta et al. (2016)	E	W	2 videos (educational and recreational)	112 students (32 F and 64 M)	9,068 clips (10 secs)	10 EA	4 levels of 4 affective states: **engagement**, frustration, confusion, boredom	Yes[b]
Zaletelj et al. (2017)	E	S	4 lecturing sessions (@25-min) in offline classroom setting	18 students	videos and kinect features	5 EA	3-level scale attention score (high, medium, low)	N/A
Monkaresi et al. (2017)	E	S	Writing task (draft-feedback-review)	22 students	1,325 video segments	Concurrent and retrospective self-report	Not engaged, engaged	N/A

Table 3 (continued)

Dataset	Type	Setting	Stimuli	Participants	Samples	Annotators	Label	Publicity
(UE-HRI) Youssef et al. (2017)	E	S	Interaction with Pepper robot	195 participants (125 M, 70 F)			Sign of Engagement Decrease (SED), Early sign of future engagement BreakDown (EBD), engagement BreakDown (BD), Temporary disengagement (TD)	Yes[c]
(BAUM-1) Zhalehpour et al. (2017)	ER	S &P	Short video clips	31 subjects (Turkish)	1,184 clips	5 EA	13 emotional and mental states, which are Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness(Sa), Surprise (Su), Boredom (Bo), Contempt (Co), Unsure (Un), Neutral (Ne), Thinking (Th), Concentrating (Con), Bothered (Bot)	Yes[d]
Hussain et al. (2018)	E	WE	Social science course on virtual learning environment (VLE)	383 students	Log file	N/A	IF (score on the assessment>= 90) OR (final results=Pass AND total number of clicks>- average clicks), then label = high engagement. Otherwise -> Low engagement	N/A
Psaltis et al. (2018)	ER	S &P	prosocial games: Path of Trust (original version and stripped-down version)	72 participants	750 videos from 15 subjects (3 seconds)	Retrospective self-reports	5 basic emotions (anger, fear, happines, sadness, surprise)	Yes[e]
Rudovic et al. (2018b)	E	S	25 min therapy session with NAO robot to learn four basic emotions: sadness, fear, anger, and happines	35 Chld.(17 from Japan, 18 from Serbia) ages 3 to 13 with autism	10s video fragment	5EA	6 engagement level [0-5] = evasive, non-compliance, indifferent, low engagement, mid engagement, high engagement	N/A
Ninaus et al. (2019)	ER	S	1) The number line estimation task, 2) watching a short clip.	122 participants	Image frames	Self-report	joy = "excited" or "inspired", activity/interest = "attentive", "active", afraid = "distressed", "scared", upset = "irritable", "hostile"	N/A
Yue et al. (2019)	ER	S &WE	MOOC course titled "Data Processing Using Python" with course 5=10 mins videos, teaching materials and quizzes.	46 participants	7224 learning performance instances	self-report and quiz score	7 emotions (Neutral, Happy, Disgust, Sad, Surprise, Fear, Anger),Eye Movement (writing, read, type), course: score	N/A

Table 3 (continued)

Dataset	Type	Setting	Stimuli	Participants	Samples	Annotators	Label	Publicity
(AffectNet) Mollahosseini et al. (2019)	ER	W	Images collected from internet	450,000 subjects	Training set: 23,901 images Validation set: 3,500 images	12 EA for 450,000 images. 2EA for 36,000 images	'8 emotion categories (neutral, happy, ,sad, surprise, fear, disgust, anger, contempt), valence and arousal (continuous)	Yes
Celiktutan et al. (2019)	E	S	HHI: dyadic interactions, HRI: triadic interactions	18 students	290 clips of HHI, 456 clips of HRI, and 746 clips in total for each data modality (276 physiological clips of HHI)	self-report	Big Five personality traits (extroversion, neuroticsm, openness, agreeableness, conscienctiousness), 10-point likert scale of engagement	Yes[f]
Youssef et al. (2019)	E	S	Interaction with Pepper robot	278 users (182 M, 96 F)	209 interactions featuring a single user, and 69 multiparty interactions	2 EA using ELAN	Sign of Engagement Decrease (SED), engaged	Yes[g]
Olivetti et al. (2019)	E	S	A virtual learning environment (A European Enterpreneurship VET Model and Assessment)	12 participants (6 F, 6 M)	3D videos	2 EA and self-report	Engagement level 1,2,3	N/A
Ashwin et al. (2020b)	E	S &P	Offline classroom	50 students	24000 posed images of 50 students, 36000 images spontaneous	slef-anotate and 2 EA	**Engaged**, boredom and neutral	N/A
Ashwin et al. (2020a)	E	S &P	Offline classroom	350 students (Indian)	2900 posed images (1450 are multiple students in a single frame), 72000 spontaneous images	30 EA	Attentive (happiness, surprise, delight, **engaged**), in-attentive (sadness, fear, disgust, boredom, sleepy, frustrated, confused).	N/A
Pabba et al. (2022)	E	S	Offline classroom	50 (31 M and 19 F) (Indian)	1193 images (30 minutes)	5 EA	Engagement level academic affective states (low:Boredom, sleepy;Medium: Yawning, frustrated, confused;High: Focused)	N/A
Duchetto et al. (2020)	E	S		227 people (122 F, 105 M. 138 adults, 89 minors)	3,106 videos (10 fpr)	3 EA Using NOVA	Engagement score :High, low, medium	N/A
Yun et al. (2020)	E	S	Interactive multi-intelligence material	20 children (Asian)	356 video/images	3 and 7 EA	Engaged, Disengaged ([High Engagement, Low Engagement], [Low Disengagement, and High Disengagement]) [17:11:5:1]	N/A

Table 3 (continued)

Dataset	Type	Setting	Stimuli	Participants	Samples	Annotators	Label	Publicity
Zhang et al. (2020)	E			47 students (28 M,19 F)	26 hours video (2 seconds image) and mouse movement	8 EA	1-5 engagement scale (but only 2 class classification Engaged, not engaged)	N/A
Liao et al. (2021)	Used Public Avilable dataset: Dataset for Affective States in E-Environments (DAiSEE)							
Li et al. (2021)	E-R	S	A virtual patient in BioWorld	61 medicalstudents	167 segments, videos (10 seconds)	self-report, 1 EA	8 clinical behaviors, 2 performances (shallow/surface, high/deep)	N/A
Bhardwaj et al. (2021)	E	S	Online class	1000 participants	Emotion scores	10 EA	0-5 scale engagement level and emotions: angry, disgust, fear, sad, surprise and neutral	Yes[h]
Goldberg et al. (2021)	E	S	offline classroom (90 mins), knowledge test	52 students (only 30 were used due to occlusions)	Videos	self-report and 6 EA using CARMA	-2 to +2 engagement scale:	
Chatterjee et al. (2021)	E	S	Dyadic conversation	16 dyads	Naturalistic conversations (15 minutes)	Self-report	Engagement level (none to very high), Engagement scale (0-100)	N/A
Youssef et al. (2021)	E	S	Interaction using Pepper robot	195 participants (70 F, 125 M)	124 interactions to feature a single user, 71 multiparty interactions (40 started as multiparty and ended as single user)	EA	Signs of User Engagement Breakdown (UEB): Breakdown, No Breakdown	N/A
Sumer et al. (2021)	E	S	Offline classroom	15 students	360 audio-visual recording	2 EA using CARMA every second	3 engagement class label (0,1,2)	N/A
Trindade et al. (2021)	ER	WE	Courses in Moodle		2752 Moodle record data from 2015-2019			N/A
Ma et al. (2021)	Used Public Avilable dataset: Dataset for Affective States in E-Environments (DAiSEE)							
Thiruthvanathan et al. (2021)		Used Public Avilable dataset: DAiSEE, iSAFE, ISED						
Altuwairqi 2021 et al. (2021b)	E	S	Writing task	42 participants	164 videos, mouse and keyboard log	Self-annotation	Strong, high, and medium engagements	N/A
Vanneste et al. (2021)	E	S	On lectures (hybrid virtual classroom)	14 students (4 F,10 M)	1031 clips (only 37-185 were annotated)	Self-report, EA	0,1,2 engagement	N/A
Hasnine et al. (2021)	E	S	Interactive lecture, lecture video taken from YouTube (28s)	11 students		N/A (concentration index (CI))	Highly-engaged, engaged, disengaged	

Table 3 (continued)

Dataset	Type	Setting	Stimuli	Participants	Samples	Annotators	Label	Publicity
Delgado et al. (2021)	E	WE	Math problem on Math-Spring.org	19 students	400 videos(18,721 frames)	3 EA	**Engaged** (looking at the screen or looking at their paper), wandering	N/A
Engwall et al. (2022)	E	S	Robot interaction (with Furhat anthropomorphic robotic head) in Wizard-of-Oz setup	33 language learners	50 audio-visual conversational videos (38 video recordings, 353 of 5s clips)	1 EA (audio recordings), 3 EA (video recordings), 9 EA (2s clips)	High and Low engagement. Clips (very dissengaged, dissengaged, neutral, engaged, very engaged)	N/A
Mehta et al. (2022)	Used Public Avilable dataset: Dataset for Affective States in E-Environments (DAiSEE)							
Dubovi et al. (2022)	ER	S	The Medication Administration Test (MAT), PANAS questionnaires	61 nursing students	Data streams, and pre-and post test context knowledge test	Self-report using PANAS	10 positive emotions and 10 negative emotions Positive and Negative Affect Scale (PANAS)(Watson et al. 1988)	N/A
Thomas et al. (2022)	Used Existed dataset[i]							
Shen et al. (2022)	Used Public Avilable dataset: JAFFE, CK+, RAF-DB							
Apicella et al. (2022)	E	S	Cognitive task (Continuous Performance Test), background music, social feedback	21 students	45 seconds acquisition EEG signals	Self-report, Performance index	High or low emotion engagement, high or low cognitive engagement	N/A

ER engagement-related dataset; *E* engagement dataset; *S* spontaneous; *P* posed; *W* in-the-Wild; *WE* web-based learning environment; *EA* external annotator;

[a] stated in the abstract but the database link is unavailable

[b] https://people.iith.ac.in/vineethnb/resources/daisee/index.html

[c] https://adasp.telecom-paris.fr/resources/2017-05-18-ue-hri/

[d] https://archive.ics.uci.edu/ml/datasets/BAUM-1

[e] https://vcl.iti.gr/masr-dataset

[f] https://www.cl.cam.ac.uk/research/rainbow/projects/mhhri/

[g] https://adasp.telecom-paris.fr/resources/2017-05-18-ue-hri/

[h] Partially

[i] IIITB Classroom Seminar dataset, IIITB Online Seminar dataset, IIITB Presentation style dataset, LectureVideoDB, ClassX, IIIT-AR-13K

Table 4 An overview of the method used in the selected articles to address RQ3

Author	Input device/modality	Input features	Estimation method	Performance metrics
Wang et al. (2010)	Thermal camera	Grayscale images pixels	**Feature extraction**: PCA, PCA + LDA, AAM, and AAM+LDA. **Classification**: KNN.**Validation**: LOOCV	Accuracy
Cocea et al. (2011)	Log file	30 log attributes	**WEKA**. 8 algorithms: 1) BNs, 2) LR, 3) simple logistic classification (SL), 4) Instance-based classification with Ibk algorithm (IBk), 5) Atrribute selected classification using J48 classifier and Best first search (ASC), 6) Bagging using REP (reduced error pruning) tree classifier (B), 7) Classification via Regression (CvR), 8) DTs	Accuracy (highest 91%)
AlZoubi et al. (2012)	3 sensors (electrocardiogram (ECG), facial electromyogram (EMG), galvanic skin response (GSR)), webcam, screen recorder.	117 features (EEG, currogator musle EMG, finger tips GSR)	**Preprocess**: low/high pass filter. **Feature extraction**: using Augsburg Biosignal Toolbox (Wagner et al.). **Classification**: PRTools 4.0 (de Ridder et al. 2017), a pattern recognition library for Matlab. 9 classifiers: 1) SVM with linear kernel (SVM1), 2). SVM with polynomial (SVM2), 3) KNN ($k = 3$). 4) KNN ($k = 5$), 5) KNN ($k = 7$). 6) NB, 7) Linear Bayes Normal Classifier (LBNC), 8) Multinomial LR, 9) C4.5 DT. **Validation**: 10-fold cross validation with 20:7 train:test ratio	Kappa statistic and F1-scores. (KNN and LBNC yielded the best detection)
S-Syun et al. (2012)	Microphone, camera, Depth sensor	Oculesic, kinesic, proxemic, vocalic, person identity cue features	**Oculesic** (gaze direction), **Kinesic** (facial expression, movement, body posture/gesture), proxemic (body posture/gesture, spatial relation), vocalic (user call), person identity cue (Spatial relation, face identification). **Feature extraction**: OpenNI library. **Binary classifications** (inattention and attention): Fuzzy-based classification algorithm (FMMNN classifier). Fuzzy min-max neural networks (FMMNN) with 7 input nodes. **Validation**: 7:3 training:test samples	Accuracy 86%

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Whitehill et al. (2014)	Camera	Facial features	**Feature extraction**: using CERT. **Binary Classification**: Boost (BF), SVM (Gabor), MLR (CERT). **Validation**: 4-fold cross-validation	2-alternative forced choice (2AFC)
Schiavo et al. (2014)	Camera	Head movement and face features	**Features extraction**: using face actions and expression recognition (Joho et al. 2011). **3-class classification**: SVM. **Validation**: LOOCV	Accuracy=73%, F-score = 63%
Woo-Han Yun et al. (2015)	Camera	55 features of face and head information	**Pre-processing**: median filtering and aggregation method (mean, median, max, min, standard deviation (STD), range, rate of zero crossings (ZCR). **4-class classification**: relevance vector classifier (RVC), a spare version of Bayesian kernel logistic regression or Gaussian process classification (GPC).	Accuracy = 78.53%, Balanced Accuracy = 70.64%
Gupta et al. (2016)	Camera	Image pixels	**Classification**: InceptionNet, C3D, LRCN.	Accuracy
Zaletelj et al. (2017)	Kinect one sensor	2D and 3D gaze point and body posture data	**3-class classification**: DT (simple and medium), KNN (coarse, medium, and weight), Bagged Trees, Subspace KNN	Accuracy = 75.3%
Monkaresi et al. (2017)	Kinect face tracker and ECG sensors (BIOPAC MP150 system)	kinect face tracker features, LBP-TOP, heart rate data	**Pre-process**: RELIEF-F for feature selection, Synthetic Minority Oversampling Technique (SMOTE) to handle the data imbalanced. **Classifications** using WEKA: Updateable NB, BN, LR, classification via cLustering, rotation forest, dagging. **Validation**: LOOCV.	AUC = 0.758 and 0.733.
Youssef et al. (2017)	No estimation method. Only proposed dataset.			
Zhalehpour et al. (2017)	Camera	Images	**Face tracking**: CHEHRA tracker. **Classification**: SVM. Accuracy: 5-class classification = 75.32%, 8-class = 65.84%	
Hussain et al. (2018)	Log file	Number of clicks and activity types	Activity types includes dataplus, forumng, glossary, oucollaborate, oucontent, resource, subpage, homepage, and URL. **Binary classification**: decision tree (DT), J48 (belongs to DT family), CART, JRIP decision rules, GBDT, NB. **Validation**: 10-fold cross validation	Accuracy, Recall, AUC, Kappa

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Psaltis et al. (2018)	Kinect face tracker	Facial expression, Body motion features, average time of responsiveness.	Feature for emotional engagement: facial expression and body motion. Feature for behavioral engagement: average time of responsiveness. **Binary classification**: unimodal ANN classifiers. **Validation**: 4-fold validation. Testing on: three primary schools.	Accuracy = 85%
Rudovic et al. (2018b)	Audiovisual sensors from NAO robot and physiological sensors to provide heart-rate, electrodermal activity, body temperature, and accelerometer data.	Face, body, physiology features, CARS, the demographic features (culture and gender)	**Pre-process**: OpenFace, OpenPose openSMILE (Eyben et al. 2013), and self-built tools for feature extraction. DeepLift for feature selection. **Regression**: personalized perception of affect network (PPA-net) whis based on ANN and clustering using t-SNE.	Intra-class correlation (ICC) = $65\% \pm 24$ (average $\pm$ SD)
Ninaus et al. (2019)	Webcam	Image frames	**Pre-process**: Microsoft's Emotion-API classifying the prevalence of the 6 basic emotions for each frame of the captured videos ('fear' and 'disgust' are excluded to enhance the quality of the data). **Classification**: SVM ensembles using "classyfire" package in R statistical environment. Questionnaires were analyzed using separate multivariate ANOVAs	Accuracy $\approx$ 64.18%
Yue et al. (2019)	Microsoft LifeCam webcam and Tobii Eye Tracker 4C	Video/images, eye movement, and click stream data.	Fine-tuning parameters by transfer learning for CNN: VGG16, InceptionResNetv2. **Classification**: CNN and LSTM. **Regression**: CART, random forest, GBDT. **Validation**: 10-fold cross validation.	Accuracy = 76.08% for facial expressions recognition, 81% for eye movement behavior. R2 metric = 0.98 ofof course performance prediction.
Mollahosseini et al. (2019)	N/A	Images	CNN (AlexNet) and SVR on Valnce and Arousal labels	RMSE, CORR, SAGR, CCC.
Celiktutan et al. (2019)	Cameras (2 static & 2 dynamic), 2 biosensors	Image, sensor data	**Binary classifictions**: SVMs. **Validation**: a double LOOCV.	
Youssef et al. (2019)	Robot's camera	Distance; head, gaze and face streams; speech;looking and listening.	**Feature extraction**: OpenFace and Pepper OKAO software. **Binary classification**: LR, DNN, GRU, LSTM. **Validation**: 3-fold cross validation	Accuracy, F1-Score, AUC
Olivetti et al. (2019)	Camera	Images (geometrical description)	**3-class classification**: SVM	The classification result was compared with the questionnaire.

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Ashwin et al. (2020b)	Camera	299x299x3 image with RGB with facial expressions, hand gestures and body postures present	**Pre-processing** = data augmentation. **Classification**: transfer learning with inception v3. **Hybrid CNN** = CNN-1 + CNN-2. CNN-1 for single student in as single image frame. CNN-2 for multiple students in a single image frame. **Validation**: 10-fold cross validation	**Posed**: accuracy = 86%, recall = 89%, precision = 91%, F1-score = 84%, AUC = 90%. **Spontaneous**: accuracy = 70%, recall = 72%, precision = 77%, F1-score = 62%, AUC = 69%
Ashwin et al. (2020a)	Camera	Images with facial expressions, hand gestures and body postures present	**Classification**: CNN with pre-trained on GoogleNet architecture(Krizhevsky et al. 2017). **Validation**: 10-fold cross validation.	Accuracy = 76%
Pabba et al. (2022)	Camera	48x48 image pixels	Add additional public dataset: BAUM-1,DAiSEE, and Yawning Detection Dataset (YawDD)[d] . **Pre-process**: face and head detection (using multi-task cascade CNN (MTCNN)), face alignment, data augmentation. **6-class classification**: CNN.	Accuracy = 76.9%
Duchetto et al. (2020)	Head camera of the robot	RGB frame-by-frame image	**Face detection**: CNN. **Regression**: LSTM. Build the model using TOGURO dataset and evaluated on UE-HRI.	AUC=0.89
Yun et al. (2020)	Camera, Kinect V2	Facial features	**Classification**: CNN with fine tunning by using a pre-trained network (VGG-3D model). **Validation**: 6-fold cross-validation, leave-one-labeler-out cross-validation (LOLOCV).	Accuracy, AUC of ROC (ROC), AUC of PRs (PRs), MCC, F1-score, balanced accuracy, specificity (true positive and negative rate).
Zhang et al. (2020)	Camera	grayscale image (100 x 100 pixel)	**Feature extraction**: adaptive weighted LGCP. **Binary classification**: fast sparse representation (AWLGCP &FSR). **Validation**: 10-fold validation. Compare: the four methods (CLBP-SRC, Gabor-SVM, active shape model-SVM, and AWLGCP &FSR).	

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Liao et al. (2021)	N/A	DAiSEE and EmotiW images	**Face detection**: MTCNN. **Pre-process**: resize images to 224×224 and pre-trained on VGGFace2. **4-class classification** and **regression**: Deep Facial Spatiotemporal Network (DFSTN) = pretrained SE-ResNet-50 (SENet) for extracting facial spatial features, and LSTM Network with Global Attention (GALN). **Validation**: 5-fold cross-validation.	Accuracy = 58.84% and MSE = 0.0422 on DAiSEE. MSE = 0.0736 on EmotiW.
Li et al. (2021)	Camera, log file	Facial features (Gaze, Pose, FAU) and 8 clinical behaviors	**Performance (correcteness)** labelling: for problem solving process (Measure cognitive engagement). **Feature extraction**: using OpenFace. Calculate mean and std of each facial features. **Feature selection**: recursive feature elimination random forest (RFE-RF). **Binary classification**: NB, KNN, DT, RF, SVM. **Validation**: 10-fold-cv for feature selection. Use students self-reports of cognitive engagement states as the ground-truth	
Bhardwaj et al. (2021)	FER-2013 dataset (image), and MES dataset	images	**Face detection**: OpenCV. **Binary classification**: CNN. First, calculating weights matrix of emotions, then calculating MES and detecting engagement.	
Goldberg et al. (2021)	3 Cameras	Eye gaze, head pose, and facial expressions.	**Feature extraction**: OpenFace. **Regression**: *Model 1*: multiple linear regression. *Model 2*: two additional linear regression. *Model 3*: add learning prerequisites.	MSE= 0.05. Pearson correclation coefficient between manual annotations' mean level and prediction models $r = .70$, $p = 0$
Chatterjee et al. (2021)	electrocardiography, skin conductance, respiration, skin temperature, Yeti X microphone, webcams	Electrocardiography, skin conductance, respiration, skin temperature signals	**Pre-process**: lowpass/highpass filter using MATLAB/Simulink. **Regression**: a binary decision tree, leastsquares boosting, and random forest implemented in MATLAB 2020b. **Validation**: LOOCV	

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Youssef et al. (2021)	Robot's camera	Distance; head, gaze and face streams; Speech;Laser	Face detection: NAOqi People Perception.**Face extraction**: OKAO Vision sofware. Imbalanced issue: undersampling "No breakdown", oversampling "Breakdown" class using SMOTE. **Binary classification**: LR, LDA, RF, and MLP. **Validation**: 5-fold cross validation.	AUC $\approx$ 0.72
Sümer et al. (2021)	Camera	Face features, head pose (without facial landmarks)	**Face detection**: RetinaFace. Multi channel settings : training Attention-Net for head pose estimation and Affect-Net for facial expression recognition CNN. **Pre-Process**: : PCA (for SVM). **3-class classification**: SVM (use majority voting), RF, MLP, LSTM with fine tunning (transfer learning) with AffectNet for facial expression and Attention-Net (300W-LP) for head pose with ResNet-50. **Tested** using different fusion strategies using RF engagement classifiers. Use of self-supervision and representation learning on unlabelled classroom data.	AUC = 0.84 (with personalization). Attention-Net better than Affect, given that the criteria for the manual annotation of engagement is not directly related to gaze direction or facial expression.
Trindade et al. (2021)	Log file	Teacher and students attributes	**WEKA**. Random Forest generated the best result.	AUC
Ma et al. (2021)	Use DAiSEE	Eye gaze, facial action unit, head pose (117 dimensions); and body pose (60 dimensions)	**Feature extraction**: OpenFace 2.0. **Pre-process**: 640x640 resolution at 10fps. **Feature Fusion**: Neural Turing Machine (NTM) architecture, which contains two basic components: a neural network controller and a memory bank. NTM workflow: read heads and write heads.	Accuracy = 60.2%
Thiruthvanathan et al. (2021)	Indian origin faces datasets DAiSEE, iSAFE, ISED	508 images from ISED and iSAFE, 5295 images from DAiSEE.	**Feature extraction**: light weight ResNet. **Classification**: ResNet classifier (CNN with 50 layers deep).	Accuracy, Precision, Recall, Sensitivity, Specificity and F1 score
Altuwairqi 2021 et al. (2021b)	Camera, mouse, keyboard behaviour	Key frame facial expressions.	Transfer learning using FER2013 and real-world affective faces (RAF). **3-class classification**: Naive Bayes (NB) classifier.	Accuracy and MSE.

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Vanneste et al. (2021)	Camera	Upper body keypoints, eye gaze direction	**Feature for individual classification**: upper body keypoints (from 2s clips), **for collective classifications**: eye gaze direction. **Classification**: i3D model (CNN based) (Carreira and Zisserman 2017). **Multilevel regression**: to investigate how the engagement cues relate to the engagement scores. Calculate the CST (collective state transition) to measure classroom engagement.	Recall and Precision. Hand-raising and note-taking are not related to students individual self-reported engagement scores.
Hasnine et al. (2021)	Camera	Video	**Face detection**: Dlib. **3-class classification**: training with FER2013, then calculate the concentration index (CI) based on eye gaze and emotion weights. CI = (Emotion Weight x Gaze Weight) / 4.5	Accuracy = 68%
Delgado et al. (2021)	Camera	Images	**Classification**: utilizing CNN family including MobileNet (Mobilenets: Efficient convolutional neural networks for mobile vision applications), VGG (Very deep convolutional network for large-scale image recognition), Xception: Deep learning with depth-wise separable convolutions.	
Engwall et al. (2022)	Cameras and microphone	Audio and visual features	**Feature extraction**: OpenFace 2.0. **Feature selection**: verbal classifications using bag-of-words representations, accoustic-based classification,video-based classification. **Engagement classification through acoustic and visual**: classification using SVM, DT, Conditional Random Fields, KNN, HMM, Gaussian model, BN, and ANN. **Engagement classification through vocal arousal**: bidirectional LSTM network Speech Emotion Recognition implementation in the Matlab Deep Learning Toolbox. *Output*: anger and happines = High, neutral = Neutral, boredom and sadness = Low. **Engagement classification through face expression**: two SVM with linear and radial basis function (RBF) as kernel.	Listener engagement classification reached 65% balanced accuracy

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Mehta et al. (2022)	Use DAiSEE and Emoti-W dataset	Images	**Pre-processing**: Dlib face detector. **4-class classification and regression**: 3D CNN with self-attention module, which enhances the discovery of new patterns in data by allowing models to learn deeper correlations between spatial or temporal dependencies between any two points in the input feature maps.	Classification accuracy = 63.59% on DAiSEE, regression MSE = 0.0347 on DAiSEE and 0.0877 = Emoti-W
Dubovi et al. (2022)	Eye tracker, EDA wearable wristband sensor, and webcam	Facial expression, eye-tracking, and EDA data	The stream data was collected and analysed using iMotion 9.0 with 7 basic emotions annotation (joy, anger, surprise, contempt, fear, sadness, and disgust). **Emotional engagement**: a Linear Mixed Effects Model (LMM) was established to estimate the self-reported changes in the PANAS self-report. **Cognitive engagement**: ANOVA was performed to assess the eye-tracking metrics differences.	
Thomas et al. (2022)	Use existed dataset[b]	Visual and verbal features	**Pre-process**: slide area and figure detection using RetinaNet, unique slide detection using Siamese network, text detection using Character-Region Awareness For Text detection (CRAFT) model. **Prediction**: pretrained with pretrained VGG-16 network. Supervised: LR with three classes (visual, verbal, or balanced). Unsupervised: clustering model with two clusters (visual, verbal). **Binary lassification**: sequential modeling using Temporal Convolutional Network (TCN) pre-trained with Micro-Macro-Motion (MIMAMO) Net model (Deng et al. 2020).	**At the segment level**: accuracy = 76%, F1-score = 0.82, MSE = 0.04. **At video level (binary classification: engaged/distracted)**: accuracy = 95%, F1-score = 0.97, MSE = 0.15

Table 4 (continued)

Author	Input device/modality	Input features	Estimation method	Performance metrics
Shen et al. (2022)	Use JAFFE, CK+, RAF-DB dataset	Images	**Pre-process**: MK-MMD to calculate the distribution distance between the extracted features. **Transfer learning**: Domain adaptation technique was used to explore the additional facial images. Imbalanced issue: undersampling, and data augmentation. **4-class classification**: lightweight attention convolutional network for face expression recognition. Soft attention module (SE) was adopted to reduce the impact of the complex background.	Accuracy = 56%
Apicella et al. (2022)	EEG	EEG Signal	**Pipeline**: Filter bank, Common Spatial Pattern, SVM. **Pre-process**: artifact removal using indpendent componenet analysis (ICA), namely Runica module of the EEGLab tool. **Feature extraction**: 12-component Filter Bank. Imbalanced problem: Stratified leace-2-trials out. **Binary classification**: SVM, Linear Discriminant Analysis (LDA), KNN, shallow ANN, DNN, CNN (pre-trained in Common spatial pattern (CSP)).	SVM achieved the highest score accuracy = 76.9% for cognitive engagement, and 76.7% for emotional engagement.

PCA principle component analysis; *LDA* linear discriminant analysis; *AAM* active appearance model; *LOOCV* leave-one-subject-out cross validation; *KNN* K-nearest neighbors; *BNs* Bayesian Nets;*LR* logistic regression; *DTs* Decision Trees; *NB* Naive Bayes; *LBP-TOP* three orthogonal planes; **GBDT** gradient boosting trees; *CART* classification and regression tree; *CARS* - childhood autism rating scale; *GRU* - gated recurrent unit; *LSTM* long-short term memory; *CNN* convolutional neural network; *KNN* K-nearest neighbors; *BNs* Bayesian Nets; *LR* logistic regression; *DTs* - Decision Trees; *NB* Naive Bayes

LOOCV leave-one-subject-out cross validation; *LR* logistic regression; *RF* - random forest; *LDA* linear discriminant analysis; *MLP* Multi-layer Perceptron; *DT* decision tree; *BN* Bayesian Network; *HMM* -Hiden Markov Models; *LSTM* long short time memory; *EDA* electrodermal activity; *MK-MMD* Kernel Maximum Mean Discrepancies

[a] https://dx.doi.org/10.21227/e1qm-hb90

[b] ClassX, LectureVideoDB, IIIT-AR-13K, IIITB Online Lecture, IIITB Classroom Lecture dataset

Acknowledgements
Not applicable.

Author contributions
SNK: conceptualization, literature selections and review, investigation, original draft preparation (writing, review and editing). SH: supervising, writing, review and editing. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by JSPS KAKENHI Grant Number 20H04294 and Photron limited.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare no competing interests.

Received: 12 July 2022 Accepted: 24 October 2022
Published online: 12 November 2022

References

Abdellaoui, B., Moumen, A., El Bouzekri El Idrissi, Y. & Remaida, A. (2020). Face detection to recognize students' emotion and their engagement: A systematic review. In: 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1–6 https://doi.org/10.1109/ICECOCS50124.2020.9314600

Abedi, A. & Khan, S.S. (2021). Improving state-of-the-art in detecting student engagement with Resnet and TCN hybrid network. In: 2021 18th Conference on Robots and Vision (CRV), pp. 151–157 https://doi.org/10.1109/CRV52889.2021.00028

ACM International Conference on Multimodal Interaction 2020: Eighth Emotion Recognition in the Wild Challenge (EmotiW) (2020). https://sites.google.com/view/emotiw2020/challenge-details

Akker, R., Hofs, D., Hondorp, H., Akker, H., Zwiers, J. & Nijholt, A. (2009). Supporting engagement and floor control in hybrid meetings, pp. 276–290 https://doi.org/10.1007/978-3-642-03320-9_26

Alarcão, S. M., & Fonseca, M. J. (2019). Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing, 10*(3), 374–393. https://doi.org/10.1109/TAFFC.2017.2714671.

Alexander, K. L., Entwisle, D. R., & Horsey, C. S. (1997). From first grade forward: Early foundations of high school dropout. *Sociology of Education, 70*(2), 87. https://doi.org/10.2307/2673158.

Altuwairqi, K., Jarraya, S. K., Allinjawi, A., & Hammami, M. (2021). Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image and Video Processing, 15*(7), 1387–1395. https://doi.org/10.1007/s11760-021-01869-7.

Altuwairqi, K., Jarraya, S. K., Allinjawi, A., & Hammami, M. (2021). A new emotion-based affective model to detect student's engagement. *Journal of King Saud University–Computer and Information Sciences, 33*(1), 99–109. https://doi.org/10.1016/j.jksuci.2018.12.008.

AlZoubi, O., D'Mello, S. K., & Calvo, R. A. (2012). Detecting naturalistic expressions of nonbasic affect using physiological signals. *IEEE Transactions on Affective Computing, 3*(3), 298–310. https://doi.org/10.1109/T-AFFC.2012.4.

Apicella, A., Arpaia, P., Frosolone, M., Improta, G., Moccaldi, N., & Pollastro, A. (2022). EEG-based measurement system for monitoring student engagement in learning 4.0. *Scientific Reports, 12*(1), 5857. https://doi.org/10.1038/s41598-022-09578-y.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems, 108,* 334–348. https://doi.org/10.1016/j.future.2020.02.075.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies, 25*(2), 1387–1415. https://doi.org/10.1007/s10639-019-10004-6.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. *User Modeling and User-Adapted Interaction, 30*(5), 759–801. https://doi.org/10.1007/s11257-019-09254-3.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*(1), 84–94. https://doi.org/10.1080/00461520.2015.1004069.

Ba, S.O. & Odobez, J.-M. (2006). Head pose tracking and focus of attention recognition algorithms in meeting rooms. In: Multimodal Technologies for Perception of Humans, pp. 345–357. Springer. https://doi.org/10.1007/978-3-540-69568-4_32

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate

Baltrusaitis, T., Robinson, P. & Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In: 2013 IEEE International Conference on Computer Vision Workshops, pp. 354–361. https://doi.org/10.1109/ICCVW.2013.54

Baltrusaitis, T., Robinson, P. & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10 https://doi.org/10.1109/WACV.2016.7477553

Baltrusaitis, T., Zadeh, A., Lim, Y.C. & Morency, L.-P. (2018). OpenFace 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66 https://doi.org/10.1109/FG.2018.00019

Baur, T., Mehlmann, G., Damian, I., Lingenfelser, F., Wagner, J., Lugrin, B., et al. (2015). Context-aware automated analysis and annotation of human–agent interactions. *ACM Transactions on Interactive Intelligent Systems, 5*(2), 1–33. https://doi.org/10.1145/2764921.

Bengio, Y. (2011). Deep learning of representations for unsupervised and transfer learning. In: Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop-Volume 27. UTLW'11, pp. 17–37. https://doi.org/10.5555/3045796.3045800

Ben-Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M. & Lim, A. (2017). UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 464–472. ACM, New York. https://doi.org/10.1145/3136755.3136814

Ben-Youssef, A., Clavel, C., & Essid, S. (2021). Early detection of user engagement breakdown in spontaneous human-humanoid interaction. *IEEE Transactions on Affective Computing, 12*(3), 776–787. https://doi.org/10.1109/TAFFC.2019.2898399.

Ben-Youssef, A., Varni, G., Essid, S., & Clavel, C. (2019). On-the-fly detection of user engagement decrease in spontaneous human-robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics, 11*(5), 815–828. https://doi.org/10.1007/s12369-019-00591-2.

Bevilacqua, D., Davidesco, I., Wan, L., Chaloner, K., Rowland, J., Ding, M., et al. (2019). Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: Evidence from a real-world classroom electroencephalography study. *Journal of Cognitive Neuroscience, 31*(3), 401–411. https://doi.org/10.1162/jocn_a_01274.

Bhardwaj, P., Gupta, P. K., Panwar, H., Siddiqui, M. K., Morales-Menendez, R., & Bhaik, A. (2021). Application of deep learning on student engagement in e-learning environments. *Computers and Electrical Engineering*. https://doi.org/10.1016/j.compeleceng.2021.107277.

Bosch, N. (2016). Detecting student engagement: Human versus machine. UMAP 2016: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, pp. 317–320. https://doi.org/10.1145/2930238.2930371

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems, 6*(2), 1–26. https://doi.org/10.1145/2946837.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2.

Brugman, H. & Russel, A. (2004). Annotating multi-media/multi-modal resources with ELAN. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon. http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf

Cao, Q., Shen, L., Xie, W., Parkhi, O.M. & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. https://doi.org/10.1109/FG.2018.00020

Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2017-January, pp. 1302–1310. https://doi.org/10.1109/CVPR.2017.143

Carlotta Olivetti, E., Violante, M. G., Vezzetti, E., Marcolin, F., & Eynard, B. (2019). Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: A preliminary approach. *Applied Sciences, 10*(1), 314. https://doi.org/10.3390/app10010314.

Carreira, J. & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A. & McOwan, P.W. (2012). Detecting engagement in HRI: An exploration of social and task-based context. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, pp. 421–428 https://doi.org/10.1109/SocialCom-PASSAT.2012.51

Castellano, G., Pereira, A., Leite, I., Paiva, A. & McOwan, P.W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. In: Proceedings of the 2009 International Conference on Multimodal Interfaces - ICMI-MLMI '09, p. 119. ACM Press, New York. https://doi.org/10.1145/1647314.1647336

Celiktutan, O., Skordos, E., & Gunes, H. (2019). Multimodal human-human-robot interactions (MHHRI) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing, 10*(4), 484–497. https://doi.org/10.1109/TAFFC.2017.2737019.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education, 96,* 42–54. https://doi.org/10.1016/J.COMPEDU.2016.02.006.

Chaouachi, M., Chalfoun, P., Jraidi, I. & Frasson, C. (2010) Affect and mental engagement: Towards adaptability for intelligent systems. In: Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23, Flairs, pp. 355–360.

Chatterjee, I., Goršič, M., Clapp, J. D., & Novak, D. (2021). Automatic estimation of interpersonal engagement during naturalistic conversation using dyadic physiological measurements. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2021.757381.

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

Chen, Y.-W. & Lin, C.-J. (2006). Combining SVMs with various feature selection strategies. In: Feature Extraction. Studies in Fuzziness and Soft Computing, vol. 207, pp. 315–324. Springer. https://doi.org/10.1007/978-3-540-35488-8_13
Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 1,* 100002. https://doi.org/10.1016/J.CAEAI.2020.100002.
Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823.
Christenson, Sandra, Reschly, Amy L., & Wylie, Cathy. (2012). *Handbook of Research on Student Engagement*. Springer. https://doi.org/10.1007/978-1-4614-2018-7.
Cocea, M., & Weibelzahl, S. (2011). Disengagement detection in online learning: Validation studies and perspectives. *IEEE Transactions on Learning Technologies, 4*(2), 114–124. https://doi.org/10.1109/TLT.2010.14.
Conti, D., Cattani, A., Di Nuovo, S. & Di Nuovo, A. (2015). A cross-cultural study of acceptance and use of robotics by future psychology practitioners. In: 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 555–560. https://doi.org/10.1109/ROMAN.2015.7333601
Darnell, D. K., & Krieg, P. A. (2019). Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures. *PloS ONE, 14*(12), 0225709. https://doi.org/10.1371/journal.pone.0225709.
De Carolis, B., D'Errico, F., Macchiarulo, N. & Palestra, G. (2019). "Engaged faces": Measuring and monitoring student engagement from face and gaze behavior. In: Proceedings–2019 IEEE/WIC/ACM International Conference on Web Intelligence Workshops, WI 2019 Companion, pp. 80–85. https://doi.org/10.1145/3358695.3361748
de Ridder, D., Tax, D. M. J., Lei, B., Xu, G., Feng, M., Zou, Y., & van der Heijden, F. (2017). *Classification Parameter Estimation and State Estimation*. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119152484.
DeepLearning.AI: Bad Machine Learning Makes Bad Science (2022). https://info.deeplearning.ai/science-plagued-by-machine-learning-mistakes-deepfakes-censor-profanity-wearable-ai-helps-impaired-walking-ensemble-models-simplified-1?ecid=ACsprvvjRjD_WkUlMQXnAK1TiHIeIgJOX2XELDoR_6xpahkNmpZLD_oxcL1fuZIAWbOw7KN2KNa5 &utm_campaign=The%20Batch &utm_medium=email &_hsmi=223142202 &_hsenc=p2ANqtz-_Jn2sqcU_uSZ2VW0RvExQAbB3YAplOltKhk6DX3uDJ1IEEfgy_XpZlKf_PpFaM-fatABYOHrJciMBEfqNa6UEA9aYcFg &utm_content=223128787 &utm_source=hs_email
Del Duchetto, F., Baxter, P., & Hanheide, M. (2020). Are you still with me? Continuous engagement assessment from a robot's point of view. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2020.00116.
Delgado, K., Origgi, J.M., Hasanpoor, T., Yu, H., Allessio, D., Arroyo, I., Lee, W., Betke, M., Woolf, B. & Bargal, S.A. (2021). Student engagement dataset. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2021-October, pp. 3621–3629. Institute of Electrical and Electronics Engineers Inc., IEEE. https://doi.org/10.1109/ICCVW54120.2021.00405
Deng, D., Chen, Z., Zhou, Y. & Shi, B. (2020). MIMAMO Net: Integrating micro- and macro-motion for video emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2621–2628. https://doi.org/10.1609/aaai.v34i03.5646
Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I. & Zafeiriou, S. (2019) RetinaFace: Single-stage dense face localisation in the wild. arXiv **abs/1905.00641**
Dewan, M.A.A., Lin, F., Wen, D., Murshed, M. & Uddin, Z. (2018). A deep learning approach to detecting engagement of online learners. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1895–1902. IEEE. https://doi.org/10.1109/SmartWorld.2018.00318
Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: A review. *Smart Learning Environments, 6*(1), 1. https://doi.org/10.1186/s40561-018-0080-z.
Dhall, A., Kaur, A., Goecke, R. & Gedeon, T. (2018). EmotiW 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 2018 on International Conference on Multimodal Interaction-ICMI '18, pp. 653–656. ACM Press. https://doi.org/10.1145/3242969.3264993
Dhall, A., Sharma, G., Goecke, R. & Gedeon, T. (2020). EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 784–789. ACM. https://doi.org/10.1145/3382507.3417973
Di Lascio, E., Gashi, S., & Santini, S. (2018). Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(3), 1–21. https://doi.org/10.1145/3264913.
D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational Psychologist, 52*(2), 104–123. https://doi.org/10.1080/00461520.2017.1281747.
D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems, 22*(4), 53–61. https://doi.org/10.1109/MIS.2007.79.
Dong, L., Di, H., Tao, L., Xu, G. & Oliver, P. (2010). Visual focus of attention recognition in the ambient kitchen. In: Asian Conference on Computer Vision, pp. 548–559. https://doi.org/10.1007/978-3-642-12297-2_53
Dresvyanskiy, D., Minker, W. & Karpov, A. (2021). Deep learning based engagement recognition in highly imbalanced data. In: Speech and Computer, pp. 166–178. https://doi.org/10.1007/978-3-030-87802-3_16
Dubovi, I. (2022). Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Computers & Education, 183,* 104495. https://doi.org/10.1016/j.compedu.2022.104495.
Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment, 29*(2), 136–151. https://doi.org/10.1177/1073191120957102.
Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System*. Palo Alto: Consulting Psychologists Press.
Engwall, O., Cumbal, R., Lopes, J., Ljung, M., & Mansson, L. (2022). Identification of low-engaged learners in robot-led second language conversations with adults. *ACM Transactions on Human-Robot Interaction, 11*(2), 1–33. https://doi.org/10.1145/3503799.

Eyben, F., Weninger, F., Gross, F. & Schuller, B. (2013). Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 835–838. ACM. https://doi.org/10.1145/2502081.2502224

Finn, J.D. & Zimmer, K.S. (2012). Student engagement: What is it? Why does it matter? In: Handbook of Research on Student Engagement, pp. 97–131. Springer. https://doi.org/10.1007/978-1-4614-2018-7_5

Fredricks, J.A. & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In: Handbook of Research on Student Engagement, pp. 763–782. Springer. https://doi.org/10.1007/978-1-4614-2018-7_37

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059.

Fuad, M. T. H., Fime, A. A., Sikder, D., Iftee, M. A. R., Rabbi, J., Al-Rakhami, M. S., et al. (2021). Recent advances in deep learning techniques for face recognition. *IEEE Access, 9,* 99112–99142. https://doi.org/10.1109/ACCESS.2021.3096136.

Gabrys, B., & Bargiela, A. (2000). General fuzzy min-max neural network for clustering and classification. *IEEE Transactions on Neural Networks, 11*(3), 769–783. https://doi.org/10.1109/72.846747.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285.

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems, 25*(1), 13–21. https://doi.org/10.1016/J.KNOSYS.2011.06.013.

Girard, J. M. (2014). CARMA: Software for continuous affect rating and media annotation. *Journal of Open Research Software.* https://doi.org/10.5334/jors.ar.

Goldberg, P., Sümer, m, Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., et al. (2021). Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review, 33*(1), 27–49. https://doi.org/10.1007/s10648-019-09514-z.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., et al. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Networks, 64,* 59–63. https://doi.org/10.1016/j.neunet.2014.09.005.

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist, 50*(1), 14–30. https://doi.org/10.1080/00461520.2014.989230.

Gudi, A., Tasli, H.E., den Uyl, T.M. & Maroulis, A. (2015). Deep learning based FACS action unit occurrence and intensity estimation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 2015-January, pp. 1–5. https://doi.org/10.1109/FG.2015.7284873

Gupta, A., D'Cunha, A., Awasthi, K. & Balasubramanian, V. (2016). DAiSEE: Towards User Engagement Recognition in the Wild **14**(8), 1–12 https://doi.org/10.48550/arXiv.1609.01885

Hall, J., Tritton, T., Rowe, A., Pipe, A., Melhuish, C., & Leonards, U. (2014). Perception of own and robot engagement in human-robot interactions and their dependence on robotics knowledge. *Robotics and Autonomous Systems, 62*(3), 392–399. https://doi.org/10.1016/j.robot.2013.09.012.

Hasnine, M. N., Bui, H. T. T., Tran, T. T. T., Nguyen, H. T., Akçapõnar, G., & Ueda, H. (2021). Students' emotion extraction and visualization for engagement detection in online learning. *Procedia Computer Science, 192,* 3423–3431. https://doi.org/10.1016/J.PROCS.2021.09.115.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning lor image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Hernandez, J., Zicheng Liu, Hulten, G., DeBarr, D., Krum, K. & Zhang, Z. (2013). Measuring the engagement level of TV viewers. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7. https://doi.org/10.1109/FG.2013.6553742

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Holmes, G., Donkin, A. & Witten, I.H. WEKA: A machine learning workbench. In: Proceedings of ANZIIS '94-Australian New Zealnd Intelligent Information Systems Conference, pp. 357–361. IEEE. https://doi.org/10.1109/ANZIIS.1994.396988

Husain, F., Dellen, B., & Torras, C. (2016). Action recognition based on efficient deep feature learning in the spatio-temporal domain. *IEEE Robotics and Automation Letters, 1*(2), 984–991. https://doi.org/10.1109/LRA.2016.2529686.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Computational Intelligence and Neuroscience.* https://doi.org/10.1155/2018/6347186.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59.

Joho, H., Staiano, J., Sebe, N., & Jose, J. M. (2011). Looking at the viewer: Analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications, 51*(2), 505–523. https://doi.org/10.1007/s11042-010-0632-x.

Jordan, M.I. (1990) Attractor dynamics and parallelism in a connectionist sequential machine. In: Artificial Neural Networks: Concept Learning, pp. 112–127.

Kapoor, S. & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. https://doi.org/10.48550/arXiv.2207.07048

Kaur, A., Mustafa, A., Mehta, L. & Dhall, A. (2018). Prediction and localization of student engagement in the wild. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8. IEEE. https://doi.org/10.1109/DICTA.2018.8615851

Keen, D. (2009). Engagement of children with autism in learning. *Australasian Journal of Special Education, 33*(2), 130–140. https://doi.org/10.1375/ajse.33.2.130.
Kipp, M. (2008). Spatiotemporal coding in ANVIL. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/521_paper.pdf
Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems, 115,* 24–35. https://doi.org/10.1016/J.DSS.2018.09.002.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90. https://doi.org/10.1145/3065386.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.
Larson, R. & Csikszentmihalyi, M. (2014). The experience sampling method. In: Flow and the Foundations of Positive Psychology, pp. 21–34. Springer. https://doi.org/10.1007/978-94-017-9088-8_2
Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: An International Journal, 46*(3), 517–528. https://doi.org/10.2224/sbp.7054.
Leite, I., McCoy, M., Ullman, D., Salomons, N. & Scassellati, B. (2015). Comparing models of disengagement in individual and group interactions. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 99–105. ACM. https://doi.org/10.1145/2696454.2696466
Li, S., Deng, W. & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2584–2593. https://doi.org/10.1109/CVPR.2017.277
Liao, J., Liang, Y., & Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence, 51*(10), 6609–6621. https://doi.org/10.1007/s10489-020-02139-8.
Libin, A. V., & Libin, E. V. (2004). Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robotherapy approach. *Proceedings of the IEEE, 92*(11), 1789–1803. https://doi.org/10.1109/JPROC.2004.835366.
Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 3045*(c), 1–1. https://doi.org/10.1109/TAFFC.2020.2981446.
Li, S., Lajoie, S. P., Zheng, J., Wu, H., & Cheng, H. (2021). Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education, 163,* 104114. https://doi.org/10.1016/J.COMPEDU.2020.104114.
Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. (2017). Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324
Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In: Face and Gesture 2011, pp. 298–305. IEEE. https://doi.org/10.1109/FG.2011.5771414
Liu, M., Shan, S., Wang, R. & Chen, X. (2014). Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756. https://doi.org/10.1109/CVPR.2014.226
Lucey, P., Cohn, J.F., Prkachin, K.M., Solomon, P.E. & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. In: Face and Gesture 2011, pp. 57–64. IEEE. https://doi.org/10.1109/FG.2011.5771462
Lufi, D., & Haimov, I. (2019). Effects of age on attention level: Changes in performance between the ages of 12 and 90. *Aging, Neuropsychology, and Cognition, 26*(6), 904–919. https://doi.org/10.1080/13825585.2018.1546820.
Lyons, M., Akamatsu, S., Kamachi, M. & Gyoba, J. (2002). Coding facial expressions with Gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205. IEEE Internet Computing. https://doi.org/10.1109/AFGR.1998.670949
Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology, 5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d.
Mason, S. J., & Weigel, A. P. (2009). A generic forecast verification framework for administrative purposes. *Monthly Weather Review, 137*(1), 331–349. https://doi.org/10.1175/2008MWR2553.1.
Ma, X., Xu, M., Dong, Y., & Sun, Z. (2021). Automatic student engagement in online learning environment based on neural turing machine. *International Journal of Information and Education Technology, 11*(3), 107–111. https://doi.org/10.18178/ijiet.2021.11.3.1497.
McDuff, D., Karlson, A., Kapoor, A., Roseway, A. & Czerwinski, M. (2012). AffectAura: An intelligent system for emotional memory. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 849–858. ACM. https://doi.org/10.1145/2207676.2208525
McNeal, K. S., Zhong, M., Soltis, N. A., Doukopoulos, L., Johnson, E. T., Courtney, S., et al. (2020). Biosensors show promise as a measure of student engagement in a large introductory biology course. *CBE-Life Sciences Education, 19*(4), 50. https://doi.org/10.1187/cbe.19-08-0158.
Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03200-4.
Minsu J., Dae-Ha, L., Jaehong, K. & Youngjo, C. (2013). Identifying principal social signals in private student-teacher interactions for robot-enhanced education. In: 2013 IEEE RO-MAN, pp. 621–626. https://doi.org/10.1109/ROMAN.2013.6628417
Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S., Paris, C. (2020). Automatic recognition of student engagement using deep learning and facial expression. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 2, pp. 273–289. Springer. https://doi.org/10.1007/978-3-030-46133-1_17

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923.

Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2017). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing, 8*(1), 15–28. https://doi.org/10.1109/TAFFC.2016.2515084.

Nakano, Y. I., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 139–148. https://doi.org/10.1145/1719970.1719990.

Ninaus, M., Greipl, S., Kiili, K., Lindstedt, A., Huber, S., Klein, E., et al. (2019). Increased emotional engagement in game-based learning—A machine learning approach on facial emotion detection data. *Computers & Education, 142,* 103641. https://doi.org/10.1016/j.compedu.2019.103641.

Noh, H., Hong, S. & Han, B. (2015). Learning deconvolution network for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1520–1528. https://doi.org/10.1109/ICCV.2015.178

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology, 61*(1), 50–69. https://doi.org/10.1002/asi.21229.

Okubo, F., Yamashita, T., Shimada, A. & Ogata, H. (2017). A neural network approach for students' performance prediction. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, pp. 598–599. ACM. https://doi.org/10.1145/3027385.3029479

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence, 2,* 100020. https://doi.org/10.1016/J.CAEAI.2021.100020.

Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*. https://doi.org/10.1111/exsy.12839.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews, 10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4.

Parkhi, O.M., Vedaldi, A. & Zisserman, A. (2015). Deep face recognition. In: Procedings of the British Machine Vision Conference 2015, pp. 1–12. https://doi.org/10.5244/C.29.41

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review, 24*(6), 1774–1784. https://doi.org/10.3758/s13423-017-1242-7.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., & Poggi, I. (2005). A model of attention and interest using gaze behavior. In: International Workshop on Intelligent Virtual Agents, pp. 229–240. Springer. https://doi.org/10.1007/11550617_20.

Peterson, P. L., Swing, S. R., Stark, K. D., & Waas, G. A. (1984). Students' cognitions and time on task during mathematics instruction. *American Educational Research Journal, 21*(3), 487–515. https://doi.org/10.2307/1162912.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review, 38*(1), 102–120. https://doi.org/10.1080/02796015.2009.12087852.

Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., & Hansen, L. K. (2017). EEG in the classroom: Synchronised neural recordings during video presentation. *Scientific Reports, 7*(1), 43916. https://doi.org/10.1038/srep43916.

Psaltis, A., Kaza, K., Stefanidis, K., Thermos, S., Apostolakis, K.C., Dimitropoulos, K. & Daras, P. (2016). Multimodal affective state recognition in serious games applications. In: IST 2016-2016 IEEE International Conference on Imaging Systems and Techniques, Proceedings, pp. 435–439. https://doi.org/10.1109/IST.2016.7738265

Psaltis, A., Apostolakis, K. C., Dimitropoulos, K., & Daras, P. (2018). Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Games, 10*(3), 292–303. https://doi.org/10.1109/TCIAIG.2017.2743341.

Qiao, W., & Bi, X. (2020). Ternary-task convolutional bidirectional neural turing machine for assessment of EEG-based cognitive workload. *Biomedical Signal Processing and Control, 57,* 101745. https://doi.org/10.1016/j.bspc.2019.101745.

Ramanarayanan, V., Leong, C.W. & Suendermann-Oeft, D. (2017a). Rushing to judgement: How do laypeople rate caller engagement in thin-slice videos of human-machine dialog? In: Interspeech 2017, pp. 2526–2530. ISCA, ISCA https://doi.org/10.21437/Interspeech.2017-1205

Ramanarayanan, V., Leong, C.W., Suendermann-Oeft, D. & Evanini, K. (2017b). Crowdsourcing ratings of caller engagement in thin-slice videos of human-machine dialog: Benefits and pitfalls. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 281–287. ACM. https://doi.org/10.1145/3136755.3136767

Ren, S., He, K., Girshick, R. & Sun, J. (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. NIPS'15, pp. 91–99. MIT Press.

Ribeiro Trindade, F., & James Ferreira, D. (2021). Student performance prediction based on a framework of teacher's features. *International Journal for Innovation Education and Research, 9*(2), 178–196. https://doi.org/10.31686/ijier.vol9.iss2.2935.

Rich, C., Ponsler, B., Holroyd, A. & Sidner, C.L. (2010). Recognizing engagement in human-robot interaction. In: 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 375–382 https://doi.org/10.1109/hri.2010.5453163

Rouast, P. V., Adam, M. T. P., & Chiong, R. (2021). Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing, 12*(2), 524–543. https://doi.org/10.1109/TAFFC.2018.2890471.

Rudovic, O., Park, H.W., Busche, J., Schuller, B., Breazeal, C. & Picard, R.W. (2019b). Personalized estimation of engagement from videos using active learning with deep reinforcement learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 217–226. https://doi.org/10.1109/CVPRW.2019.00031

Rudovic, O., Utsumi, Y., Lee, J., Hernandez, J., Ferrer, E.C., Schuller, B. & Picard, R.W. (2018a). CultureNet: A deep learning approach for engagement intensity estimation from face images of children with autism. In: IEEE International Conference on Intelligent Robots and Systems, pp. 339–346. https://doi.org/10.1109/IROS.2018.8594177

Rudovic, O., Zhang, M., Schuller, B. & Picard, R. (2019a). Multi-modal active learning from human data: A deep reinforcement learning approach. In: 2019 International Conference on Multimodal Interaction, pp. 6–15. ACM. https://doi.org/10.1145/3340555.3353742

Rudovic, O., Lee, J., Dai, M., Schuller, B., & Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*. https://doi.org/10.1126/scirobotics.aao6760.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W. & Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: HRI 2011-Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction, pp. 305–311. https://doi.org/10.1145/1957656.1957781

Sayash Kapoor, Priyanka Nanayakkara, Kenny Peng, Hien Pham. & Arvind Narayanan. (2022). The reproducibility crisis in ML-based science https://sites.google.com/princeton.edu/rep-workshop?utm_campaign=The%20Batch &utm_medium=email &_hsmi=223142202 &_hsenc=p2ANqtz-9bv16UMU819WtwyR5st61wc5IsAY27TZ3DBYTsGNcHzkmoYckmHvNSrW6AxtVgRZBSlu0w8dh_5h6c9GEY7Bil_my3sQ &utm_content=223128787 &utm_source=hs_email

Schiavo, G., Cappelletti, A., & Zancanaro, M. (2014). Engagement recognition using easily detectable behavioral cues. *Intelligenza Artificiale, 8*(2), 197–210. https://doi.org/10.3233/IA-140073.

Schmidt, A. & Kasiński, A. (2007). The Performance of the Haar Cascade Classifiers Applied to the Face and Eyes Detection, pp. 816–823. https://doi.org/10.1007/978-3-540-75175-5_101

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211–232. https://doi.org/10.1016/0010-0285(89)90008-X.

Schroff, F., Kalenichenko, D. & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Schuller, B. (2015). Deep learning our everyday emotions. Advances in neural networks: Computational and theoretical issues, pp. 339–346. https://doi.org/10.1007/978-3-319-18164-6_33

Sharkawy, Abdel-Nasser. (2020). Principle of neural network and its main types: Review. *Journal of Advances in Applied & Computational Mathematics, 7,* 8–19. https://doi.org/10.15377/2409-5761.2020.07.2.

Sharkawy, Abdel-Nasser. (2021). A survey on applications of human-robot interaction. *Sensors & Transducers Journal, 251*(4), 19–27.

Shen, J., Yang, H., Li, J., & Cheng, Z. (2022). Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimedia Systems, 28*(2), 469–478. https://doi.org/10.1007/s00530-021-00854-x.

Simonyan, K. & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, pp. 1–14.

Simpson, P. K. (1992). Fuzzy min-max neural networks. I. Classification. *IEEE Transactions on Neural Networks, 3*(5), 776–786. https://doi.org/10.1109/72.159066.

Sumer, O., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2021). Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3127692.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Thiruthuvanathan, M., Krishnan, B., & Rangaswamy, M. A. D. (2021). Engagement detection through facial emotional recognition using a shallow residual convolutional neural networks. *International Journal of Intelligent Engineering and Systems, 14,* 236–247.

Thomas, C., Puneeth Sarma, K. A. V., Swaroop Gajula, S., & Jayagopi, D. B. (2022). Automatic prediction of presentation style and student engagement from videos. *Computers and Education: Artificial Intelligence, 3,* 100079. https://doi.org/10.1016/j.caeai.2022.100079.

Thong Huynh, V., Kim, S.-H., Lee, G.-S. & Yang, H.-J. (2019). Engagement intensity prediction with facial behavior features. In: 2019 International Conference on Multimodal Interaction, pp. 567–571. ACM. https://doi.org/10.1145/3340555.3355714

Tincani, M., Travers, J., & Boutot, A. (2009). Race, culture, and autism spectrum disorder: understanding the role of diversity in successful educational interventions. *Research and Practice for Persons with Severe Disabilities, 34*(3–4), 81–90. https://doi.org/10.2511/rpsd.34.3-4.81.

Tingfan, Wu., Butko, N. J., Ruvolo, P., Whitehill, J., Bartlett, M. S., & Movellan, J. R. (2012). Multilayer architectures for facial action unit recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42*(4), 1027–1038. https://doi.org/10.1109/TSMCB.2012.2195170.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), vol. 2015 Inter, pp. 4489–4497 https://doi.org/10.1109/ICCV.2015.510

Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., Depaepe, F., & Noortgate, W. V. D. (2021). Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement. *Mathematics, 9*(3), 1–20. https://doi.org/10.3390/math9030287.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 6000–6010. Curran Associates Inc. https://dl.acm.org/doi/10.5555/3295222.3295349

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. 511–518. https://doi.org/10.1109/CVPR.2001.990517

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision, 57*(2), 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb.
Voit, M. & Stiefelhagen, R. (2008). Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: Proceedings of the 10th International Conference on Multimodal Interfaces - IMCI '08, p. 173. ACM Press. https://doi.org/10.1145/1452392.1452425
Wagner, J., Jonghwa Kim, Andre, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: 2005 IEEE International Conference on Multimedia and Expo, pp. 940–943. IEEE. https://doi.org/10.1109/ICME.2005.1521579
Wang, Y., Kotha, A., Hong, P.H. & Qiu, M. (2020). Automated student engagement monitoring and evaluation during learning in the wild. In: Proceedings-2020 7th IEEE International Conference on Cyber Security and Cloud Computing and 2020 6th IEEE International Conference on Edge Computing and Scalable Cloud, CSCloud-EdgeCom 2020, pp. 270–275. https://doi.org/10.1109/CSCloud-EdgeCom49738.2020.00054
Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing, 429,* 215–244. https://doi.org/10.1016/J.NEUCOM.2020.10.081.
Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., et al. (2010). A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia, 12*(7), 682–691. https://doi.org/10.1109/TMM.2010.2060716.
Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063.
Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing, 5*(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163.
Winata, G.I., Kampman, O.P. & Fung, P. (2018). Attention-based LSTM for psychological stress detection from spoken language using distant supervision. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6204–6208. https://doi.org/10.1109/ICASSP.2018.8461990
Winne, P. H., & Perry, N. E. (2000). Measuring Self-Regulated Learning. Handbook of Self-Regulation, pp. 531–566. https://doi.org/10.1016/B978-012109890-2/50045-7.
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. (2006) ELAN: A professional framework for multimodality research. In: LREC.
Witten, Ian, & Frank, Eibe. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann.
Wolters, C.A. & Taylor, D.J. (2012). A self-regulated learning perspective on student engagement. In: Handbook of Research on Student Engagement, pp. 635–651. Springer. https://doi.org/10.1007/978-1-4614-2018-7_30
Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P. & Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV), vol. 2015 Inter, pp. 3756–3764. https://doi.org/10.1109/ICCV.2015.428
Wu, J., Yang, B., Wang, Y. & Hattori, G. (2020). Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 777–783. ACM. https://doi.org/10.1145/3382507.3417959
Xie, K., Heddy, B. C., & Greene, B. A. (2019). Affordances of using mobile technology to support experience-sampling method in examining college students' engagement. *Computers & Education, 128,* 183–198. https://doi.org/10.1016/j.compedu.2018.09.020.
Yang, D., Alsadoon, A., Prasad, P.W.C., Singh, A.K. & Elchouemi, A. (2018). An emotion recognition model based on facial recognition in virtual learning environment. In: Procedia Computer Science, vol. 125, pp. 2–10. https://doi.org/10.1016/j.procs.2017.12.003
You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education, 29,* 23–30. https://doi.org/10.1016/J.IHEDUC.2015.11.003.
Yue, J., Tian, F., Chao, K.-M., Shah, N., Li, L., Chen, Y., & Zheng, Q. (2019). Recognizing multidimensional engagement of e-learners based on multi-channel data in e-learning environment. *IEEE Access, 7,* 149554–149567. https://doi.org/10.1109/ACCESS.2019.2947091.
Yun, S.-S., Choi, M.-T., Kim, M., & Song, J.-B. (2012). Intention reading from a Fuzzy-based human engagement model and behavioural features. *International Journal of Advanced Robotic Systems*. https://doi.org/10.5772/50648.
Yun, W.-H., Lee, D., Park, C., & Kim, J. (2015). Automatic engagement level estimation of kids in a learning environment. *International Journal of Machine Learning and Computing, 5*(2), 148–152. https://doi.org/10.7763/IJMLC.2015.V5.499.
Yun, W. H., Lee, D., Park, C., Kim, J., & Kim, J. (2020). Automatic recognition of children engagement from facial video using convolutional neural networks. *IEEE Transactions on Affective Computing, 11*(4), 696–707. https://doi.org/10.1109/TAFFC.2018.2834350.
Zadeh, A., Lim, Y.C., Baltrusaitis, T. & Morency, L.-P. (2017). Convolutional experts constrained local model for 3D facial landmark detection. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), vol. 2018-January, pp. 2519–2528. https://doi.org/10.1109/ICCVW.2017.296
Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing, 2017*(1), 80. https://doi.org/10.1186/s13640-017-0228-8.
Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2017). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing, 8*(3), 300–313. https://doi.org/10.1109/TAFFC.2016.2553038.
Zhang, Z., Hu, Y., Liu, M. & Huang, T. (2007). Head pose estimation in seminar room using multi view face detectors, pp. 299–304 https://doi.org/10.1007/978-3-540-69568-4_27
Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y. & Li, J. (2019). An novel end-to-end network for automatic student engagement recognition. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 342–345. https://doi.org/10.1109/ICEIEC.2019.8784507

Zhang, Z., Li, Z., Liu, H., Cao, T., & Liu, S. (2020). Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research, 58*(1), 63–86. https://doi.org/10.1177/0735633119825575.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342.

Zhao, S., Wang, S., Soleymani, M., Joshi, D., & Ji, Q. (2019). Affective computing for large-scale heterogeneous multimedia data. *ACM Transactions on Multimedia Computing, Communications, and Applications, 15*(3s), 1–32. https://doi.org/10.1145/3363560.

Zheng, X., Hasegawa, S., Tran, M.-T., Ota, K. & Unoki, T. (2021). Estimation of learners' engagement using face and body features by transfer learning, pp. 541–552. https://doi.org/10.1007/978-3-030-77772-2_36

Zhu, B., Lan, X., Guo, X., Barner, K.E. & Boncelet, C. (2020). Multi-rate attention based gru model for engagement prediction. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 841–848. ACM. https://doi.org/10.1145/3382507.3417965

Zhu, X., Lei, Z., Liu, X., Shi, H. & Li, S.Z. (2016). Face alignment across large poses: A 3D solution. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2016-December, pp. 146–155. https://doi.org/10.1109/CVPR.2016.23

Publisher's Note