Doctoral Dissertation

Automatic Recognition and Analysis of Learners' Engagement in Distance Learning Process

Shofiyati Nur Karimah

Supervisor: Shinobu Hasegawa

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information Science]
March, 2024

Abstract

Engagement is an essential component of the learning processes associated with positive learning outcomes. Measuring learner engagement in learning processes is important for providing insights for enhancing learning activities. Because the learning paradigm has shifted to enable more distance learning practices, machine learning-based automatic engagement estimation methods have been proposed as a new way to measure learner engagement. Nevertheless, most existing methods are built standalone and have yet to be integrated into actual distance learning practice. Furthermore, implementing automatic engagement estimation should ensure technological and ethical impact responsibilities.

This study aims to provide an intermediary knowledge and solution to analyse learners' engagement in the distance learning process by addressing the main research question: "How do educators or education institutions safely apply automatic engagement estimation in their distance learning process?" A systematic review is conducted to gain basic knowledge of the current trend of automatic engagement estimation in the literature to achieve this goal. The engagement types, datasets, and methods are defined and theoretically investigated. Secondly, the technical investigation to understand the basic requirement for automatic engagement estimation is done by building an engagement estimation module using deep learning methods. We introduce a design principle for end-to-end integration of real-time automatic engagement estimation in distance learning practice. Thirdly, we introduce a design principle for the ethical implementation of automatic engagement estimation so that the technology can benefit actual distance learning in practice.

From the literature review, we found that clearer engagement definitions and cues are crucial for developing an applicable automatic engagement estimation. However, there is no clear taxonomy to define engagement, especially for distance learning implementations. Therefore, we introduced a taxonomy of engagement definitions and cues, categorized the engagement datasets, and conducted method categorization, which mainly utilised machine learning-based methods. The combination of a clear definition of engagement and suitable machine learning methods allows learners' engagement during learning activities to be measured automatically, including in human-human interactions, human-computer interactions, and human-robot interactions.

Two deep learning models were experimented with, i.e., long short-term memory (LSTM) and convolutional neural network (CNN), and a publicly available engagement dataset. However, we found that classic machine learning

would be the best practice, especially for real-time engagement estimation, while LSTM is less feasible for practical implementation compared to CNN from a runtime perspective. Furthermore, a framework for real-time automatic engagement estimation is proposed for implementation in distance learning practices. Furthermore, we introduce system designs and prototypes for both an asynchronous and a synchronous setting.

We propose the design of RAMALAN, a real-time engagement assessment for asynchronous distance learning, and MeetmEE (pronounced as 'meet me'), a real-time video conference integrated with automatic engagement estimation for synchronous distance learning. The MeetmEE prototype was deployed in a pilot experiment to evaluate the MeetmEE system design. A total of 20 participants joined the experiment in a one-hour meeting session with the author via MeetmEE online either as educators ($n = 13; 65\%$) or learners ($n = 7; 35\%$) with 60%. The participants completed two survey forms (Forms A and B) based on their roles in their affiliations. The experiment results of Form A demonstrate that most of the responses were very positive to the automatic engagement estimation concept, represented in MeetmEE. MeetmEE is favourable for 70% of the participants, where, for educators, this technology will motivate them to improve their teaching strategies and give support to their students, while students can measure their own engagement as well. Furthermore, the results of Form B showed a positive evaluation, demonstrating that MeetmEE is sufficient, particularly in scales of stimulation, attractiveness, perspicuity, and novelty. However, MeetmEE is perceived as relatively low in terms of dependability and efficiency.

Finally, the user evaluation results are considered to construct the design principle of ethical implementation. The automatic engagement estimation implementation's design principle incorporates technical and operational measures. While the current automatic engagement estimation studies focused on only the ICT point of view instead of the feasibility of the actual education process, the development of an engagement estimation design principle incorporated with its real-time application in the distance learning process is a part of the originality of this research. We believed that this contribution would be beneficial in designing a broader distance learning framework where the learners' internal state and affective factors are considered.

Index Terms: Distance learning, automatic engagement estimation, emotional engagement, machine learning, WebRTC, design principle, ethical impact.

Acknowledgements

First, I am grateful to the Almighty God, Allah, for His wisdom that has been bestowed upon me and made this study possible.

Secondly, I would like to express my sincere gratitude to my supervisor, Prof. Shinobu Hasegawa, for allowing me to join his lab and work in his group and for his continuous support and encouragement, especially during this doctoral program. His insights, motivation, and guidance helped me immensely, especially in understanding how to conduct research, extract the meaning, and write research articles.

Thirdly, I would like to express my gratitude to my supervising committee, Prof. Kazunori Kotani as secondary supervisor, Prof. Shogo Okada as minor research advisor, as well as Prof. Ikeda Kokolo and Prof. Masayuki Murakami as the rest of my thesis committee. Their insightful comments and valuable suggestions inspired me to present the research works clearly and logically.

I also take this opportunity to record my sincere thanks to my fellow lab-mates and friends for their support and warm friendship throughout my research period. I would also dedicate this work to my dearest husband, Günter Ellrott, and my family for their long-lasting prayer, unconditional love, warm encouragement, and support throughout my life.

In addition, I would like to express my most profound appreciation to Doctor Research Fellow for providing me the financial support and research topics to carry out my research, study, and life in Japan.

Shofiyati Nur Karimah
Ishikawa, Japan

Contents

List of Figures

List of Tables

List of Abbreviations

AU	Action Unit
COVID-19	Coronavirus 2019 disease
HCI	Human-Computer Interaction
HHI	Human-Human Interaction
HRI	Human-Robot Interaction
ICT	Information and Communications Technology
ISA	Information Security Awareness
LMS	Learning management system
LR	Logistic Regression
LSTM	Long-Short Term Memory
MOOCs	Massive Open Online Courses
MSE	Mean Squared Error
PC	Personal Computer

Chapter 1

Introduction

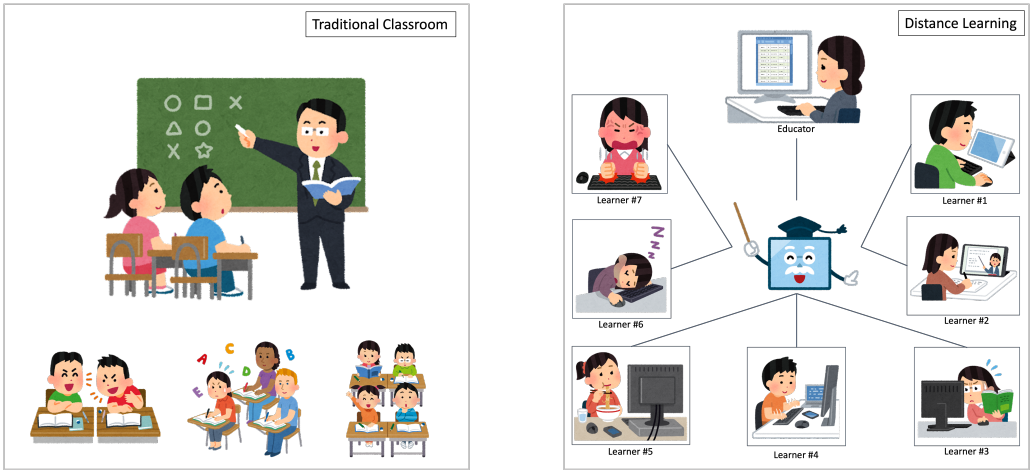
1.1 Background

Due to rapid developments in information and communications technology (ICT) and the spread of the COVID-19 pandemic, the learning model shifted from a traditional classroom to distance learning. Distance learning, including self-directed and online learning, has become a major learning setting. Distance learning enables learners to have more freedom to participate via synchronous learning (such as live communication) or asynchronous learning (such as a learning portal through a learning management system (LMS)). This freedom affects learner interaction, engagement, motivation, and learning assessments.

Learner engagement is an inner state associated with a learning process and positively correlated with academic achievement [122], and higher engagement levels lead to better learning outcomes [164]. A good engagement state is associated with curiosity, interest, optimism, and passion, which enhances motivation to continue learning and pursue achievement [78]. Therefore, understanding learners' engagement is an essential component in a learning process that: (1) increases productivity and learning; (2) provides insight for enhancing learner-educator, learner-learning material, and learner-learner interactions; (3) yields insights to improve course content delivery and lecture planning; (4) provides personalized support to learners; (5) reduces the dropout rates in online courses [7, 146, 223, 60, 197].

Unlike face-to-face offline learning, assessing learners' engagement in the distance learning process is more challenging. In traditional classrooms, educators can directly recognize how to engage their students in the class. For example, some learners show active participation, frustration, or distraction during learning (Figure 1.1). In contrast, learner engagement is more difficult

to estimate due to a limitation of learner-educator interaction in distance learning settings. Therefore, educators are called to seek different ways to assess and monitor affective behaviours in real time, including boredom, frustration, confusion, and engagement. Real-time engagement assessment benefits educators to adjust their teaching strategy the way they do in a traditional classroom, e.g., by suggesting some useful reading materials or changing the course contents [229].



(a) (b)

Figure 1.1: Problem statement of traditional classroom vs distance learning.

The recent success of artificial intelligence research and applications that support classic machine learning and deep neural networks have led to promising research on automatic engagement estimation [90, 43]. Several automatic engagement estimation methods have been proposed in recent years [111]. Among them, computer vision-based techniques are the most popular methods because nonverbal behaviours (including head motion, eye gaze, and body pose) play key roles in determining engagement levels [27]. 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 [64].

Moreover, physiological information-based methods have also received considerable attention in automatic engagement estimation research. The development of cost-effective bio-signal hardware, such as electroencephalogram (EEG), electrocardiogram (ECG), facial electromyogram (fEMG), and galvanic skin response (GSR), have provided simple and easy-to-use solutions [6].

Furthermore, physiological signals support personalized analyses, which is pertinent for learners with special needs, such as those with autism [175].

Despite the massive development of automatic engagement estimation [110, 111], there is no research on using engagement estimation modules in distance learning practice. The current automatic engagement estimation studies focused only on the ICT point of view instead of the implementation of actual distance learning processes. This challenge motivates this thesis to focus on the framework and requirements to optimize the benefit of automatic engagement estimation in accordance with distance learning characteristics.

The rest of this chapter states the motivation of the study with regard to automatic engagement estimation development to address distance learning problems. The motivations include research problems on defining engagement to be measured, dataset and methods used for the automatic engagement estimation module, and an overview of existing solutions and remaining issues. Subsequently, the significance and objective of this research will be defined in the succeeding section, followed by a section that describes distance learning characteristics. Finally, the organization of the thesis is shown at the end of this chapter.

1.2 Research Motivation

Studies on automatic engagement estimation methods have been introduced by multidisciplinary fields, including human-computer interactions (HCIs), human-robot interactions (HRIs), and embodied conversational agents (ECAs). Some studies have different methods and perspectives in defining the engagement to be measured. Therefore, there is bias in defining engagement in literature, particularly for distance learning purposes.

Besides, machine learning-based automatic engagement estimation is the existing solution for analysing learner engagement automatically [111]. For developing automatic engagement estimation methods, adequately labeled data and a sufficient amount of data that includes as many generalized variables as possible are important criteria, such as using publicly available or self-collected datasets. Publicly available datasets are open, freely downloadable, and may have some terms and conditions, such as use only in research contexts or with author consent. 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. However, to our knowledge, no research has reviewed the datasets and methods used in literature to develop an automatic engagement estimation in education. Therefore, the systematic review will cover the dataset and methods used in

the literature to develop an automatic engagement estimation module.

Furthermore, there is a gap between ICT development (i.e., automatic engagement estimation studies) and educational practices (i.e., distance learning studies). Various types of distance lectures have been used in the past distance education, such as open universities and MOOCs. Automatic engagement estimation is aimed at analysing learner engagement when traditional face-to-face methods are transferred to distance education. However, the current engagement estimation research focuses more on computer science.

For example, school principals who do not have the support of a data specialist to assist them and their teachers in adhering to demands to use data being exerted upon them by policy created above the school level [142]. In other words, although automatic engagement estimation has been introduced, it cannot immediately impact the distance learning process, especially with the lack of a technologically savvy environment for educators and education managers to interpret the report. Besides, involving a third party potentially raises a budget issue, harms data privacy, and can be ethically abusive. Therefore, the problem arising from automatic engagement estimation needs to be considered.

This issue motivates this study to bridge the computer science results with educational practice. It is wrapped in the main research question: "How do educators or education institutions safely apply automatic engagement estimation in their distance learning process?"

1.3 Research Objectives

The ultimate goal of this study is to provide an intermediary solution to analyse learners' engagement in the distance learning process, which aims to address the main research question as shown in Figure 1.2.

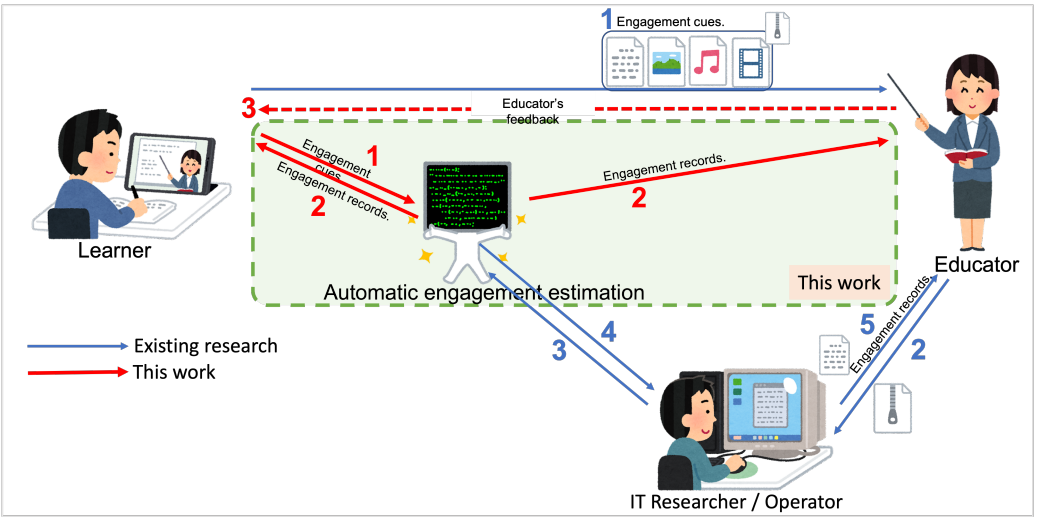


Figure 1.2: The objective of this work.

We propose a framework for enhancing distance learning performance using an automatic engagement estimation module to recognize learner engagement. To reach this objective, the main research question is broken down into three research questions:

1. What requirements did literature develop for automatic engagement estimation?
   (a) How should the type of engagement to be measured be defined?
   (b) What datasets are suitable for developing automatic engagement estimation methods?
   (c) What automatic engagement estimation methods have been developed in the literature?
2. How to develop real-time engagement estimation tools for distance learning practice?
3. How to implement automatic engagement estimation in distance learning while taking into account distance learning characteristics and ethical impact?

First, the basic knowledge of developing automatic engagement estimation in the literature will be studied. The engagement types, datasets, and methods are defined and investigated to address the RQ1. Secondly, we introduce a framework to show an end-to-end real-time automatic engagement estimation

integration based on the proposed mechanism to address RQ2. We propose MeetmEE (pronounced as 'meet me'), a real-time video conference integrated with automatic engagement estimation for enhanced distance learning. Finally, to address the RQ3, MeetmEE is deployed in a pilot experiment to evaluate the MeetmEE system design, where the user evaluation results are considered to construct the design principle of ethical implementation. We introduce the design principle of the automatic engagement estimation implementation that incorporates both technical and operational measures.

1.4 Organization of the Thesis

This thesis is comprised of six chapters. Apart from this introduction chapter, the organization of the remaining chapters, from Chapter 2 to Chapter 6, is as follows.

Chapter 1 (this chapter) describes the research's background, motivation, and objective. The thesis outline is also presented.

Chapter 2 presents a systematic review to understand the fundamental concepts, techniques, and algorithms used in engagement estimation research.

Chapter 3 investigates technical requirements to build an automatic engagement estimation module and presents our proposed real-time automatic engagement estimation framework.

Chapter 4 presents the RAMALAN and MeetmEE system design as the implementation of automatic engagement estimation in distance learning practice.

Chapter 5 discusses the ethical issues of implementing automatic engagement estimation in actual distance learning and proposes the design principle of an automatic engagement estimation system in a synchronous distance learning Practice.

Chapter 6 concludes the thesis, mentions the contribution and limitation of this thesis, and provides some insights for future works.

Chapter 2

Literature Review

This chapter is an update and abridged version of the following publications:

1. S. N. Karimah and S. Hasegawa, "Automatic Engagement Recognition for Distance Learning Systems: A Literature Study of Engagement Datasets and Methods," in International Conference on Human-Computer Interaction, 2021, pp. 264–276, doi: 10.1007/978-3-030-78114-9_19,

2. S. N. Karimah and Shinobu Hasegawa, "Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. Smart Learning Environments, 9(1):31, 11 2022, doi: 10.1186/s40561-022-00212-y.

2.1 Chapter Introduction

To address RQ1, this chapter reviews the background knowledge related to engagement definitions, datasets, and methods based on 47 articles selected in the systematic 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 [158]. A literature search was carried out based on the PRISMA flow diagram, with modifications made to the 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 2.1 shows the modified PRISMA flowchart used to select articles in this review, where the detail of each phase is explained in Appendix A.

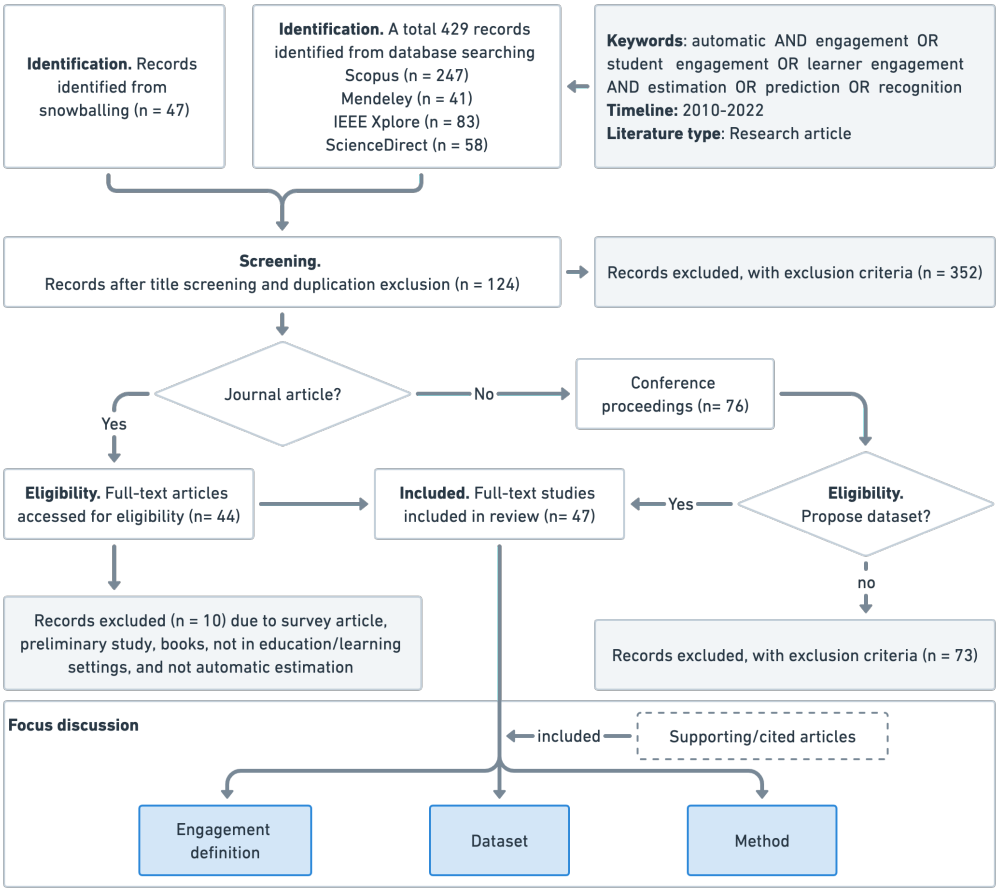


Figure 2.1: Illustration of the article selection process.

2.2 Engagement Definition for Automatic Engagement Estimation

RQ1.1: 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 [49, 113] and stimuli, such as human-computer interactions (HCIs), human-robot interactions (HRIs), and embodied conversational agents (ECAs), including human-human interactions (HHIs).

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. The second approach defines engagement as an interaction quality metric. Moreover, researchers who focused on ECAs [161] and intelligent tutor systems (ITSs) [65] 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 [223, 146, 197].

This inconsistent definition of engagement in the literature due to the lack of consensus and taxonomy for learning engagement [234] may confuse 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 (Figure 2.2). As a baseline, we follow the definition of engagement in education and learning environments proposed by Fredricks et al. ([78]), which has been widely used in engagement research [227, 77, 89, 231, 16].

Engagement is associated with internal states constructed by various cues and may not be visually apparent. Fredricks et al. ([78]) divided engagement into three categories: behavioural, emotional, and cognitive engagement. However, in this definition of engagement, the components to construct each type of engagement overlap considerably, as shown in Figure 2.2.

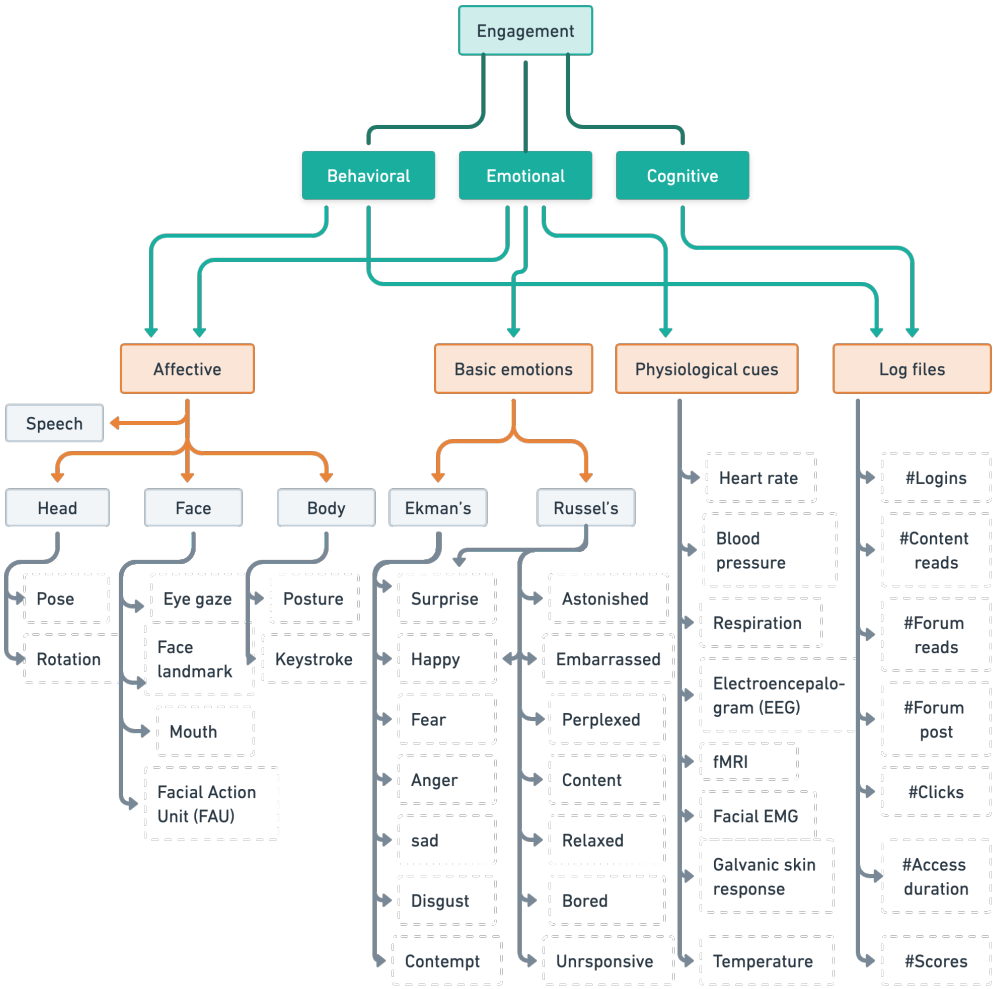


Figure 2.2: Proposed taxonomy to define engagement definition based on Fredrick's engagement category [78].

Behavioural engagement describes learners' participation in learning and tasks [78]. In classroom settings, behavioural engagement is shown by actively participating in class, such as by asking questions or displaying attention and concentration [197]. **Emotional** engagement refers to learners' affective reactions in the classroom or during learning, including interest, boredom, happiness, sadness, and anxiety [78]. **Cognitive** engagement, also referred to as self-regulation, incorporates learners' psychological investment in learning, including 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 [72]) to the learning materials for **emotional** engagement; and metacognitive strategies, namely, how learners set goals, plan, and organize their study efforts, for **cognitive** engagement [78].

In developing automatic engagement estimation methods, these components can be obtained with several modalities (Table 1), such as **log files**, which include information related to learner performance, reaction times, and errors [41, 155, 233]; **affective cues**, including face and body analyses from video/images [223, 32, 31]; and **physiological cues**, such as galvanic skin responses [63, 140], electroencephalograms (EEGs) [165, 29], heart rates [52, 146], and combinations of these cues [64].

The engagement level can be determined by grouping emotions according to Ekman's basic emotions [72] or Russel's model [179]. For example, Altuwairqi et al. ([9]) 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 the 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. ([37]) 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.

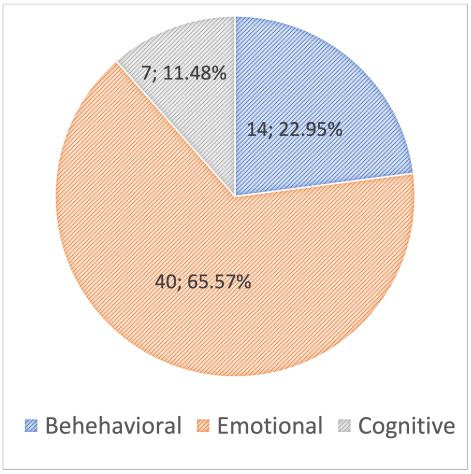
Consulting the taxonomy, we then reviewed the definition of engagement with a two-step approach. First, we examined the modalities used in each article and how the engagement level was determined. The articles included three common engagement modalities: affective cues (including audio and visual), physiological 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). As previously discussed, one engagement cue does not exclusively correspond to one engagement type.

For example, Apicella et al. ([13]) 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 Rus-

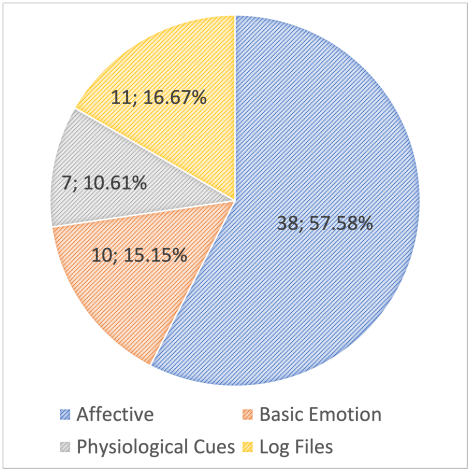
sel's four quadrants, were used to estimate emotional engagement, while a cognitive task (Continuous Performance Test) was used to estimate cognitive engagement.

Moreover, Goldberg et al. ([85]) 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 was 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 (Figures 2.3a and 2.3b) (See Appendix Table 1). In this thesis, we follow the majority of studies to use visual-based analysis from video to estimate emotional engagement.



(a) Engagement type.



(b) Engagement cues.

Figure 2.3: Pie chart of the engagement types estimated and cues measured in the selected articles.

2.3 Engagement Dataset to Build Engagement Estimation Methods

RQ1.2: What datasets are suitable for developing automatic engagement estimation methods?

Adequately labeled data and sufficient 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 using publicly available or self-collected datasets. Publicly available datasets are open, freely downloadable, and may have some terms and conditions, such as use only in research contexts or with author consent. 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 labeled 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 labeling process. However, as shown in the taxonomy of engagement estimation (Figure 2.2), 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 used in the automatic engagement estimation literature even though they have no straightforward engagement labels as **engagement-related datasets** and datasets with 'engagement' as a 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] [219], 2) BAUM-1[2] [240], 3) the MASR dataset[3], which is used in [167] but was proposed in [167], and 4) AffectNet [144]. The public **engagement datasets** include: 1) DAiSEE[4]

[1] A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference

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

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

[4] Dataset for Affective States in E-Environment `https://people.iith.ac.in/vineethnb/resources/daisee/index.html`

[91], 2) UE-HRI[5] [26], 3) MHHRI[6] [40], and 4) MES dataset[7] (Table 2.1).

DAiSEE is one of the most popular publicly available engagement datasets used in the literature [157, 127, 136, 202, 141]. 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 [61] and 2020 [62], are accessible for academic research [3].

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 videos' quality (e.g., illumination, background noise, and occlusion) varies. Although in-the-wild data have considerable variations, they are believed to be the closest to real-world conditions [91, 61, 62].

Despite the ease and amount of available data, DAiSEE, EmotiW, and other publicly available datasets, most of the datasets were collected with participants of a certain ethnicity, which may not be appropriate for all target subjects. Moreover, large variations may make' in-the-wild' data difficult to process. Therefore, most engagement studies build custom engagement datasets that address the requirements of their model or system (see Appendix A.2.2).

In contrast to publicly available datasets, the non-public (also referred to as custom) engagement datasets (Table 2, 3) were collected more than engagement-related datasets (Table 4). We found that a different way of interpreting and defining engagement is a reasonable decision to have an engagement label explicitly in the dataset. 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 described in Section 2.4. Transfer learning generally involves using a pre-trained neural network on a large dataset to extract features for tasks with smaller datasets. Some image datasets used for transfer learning include FER-2013 [87], VGGFace [159], VGGFace2 [35], FaceNet [188], AffectNet [144], 300W-LP and AFLW2000

[5]User Engagement in Spontaneous Human-Robot Interactions `https://adasp.telecom-paris.fr/resources/2017-05-18-ue-hri/`

[6]Multimodal Human-Human-Robot Interactions Dataset for Studying Personality and Engagement `https://www.cl.cam.ac.uk/research/rainbow/projects/mhhri/`

[7]Partially `https://github.com/Harsh9524/MES-Dataset/blob/main/MES_dataset.csv`

Table 2.1: Publicly available engagement-related and engagement dataset.

Dataset	Setting	Stimuli	Participants	Samples	Annotators	Label
USTC-NVIE [219]	S&P	3-4 mins emotional and 1-2 mins neutral videos from internet	215 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**.
BAUM-1 ([240])	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)
MASR [166]	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, happiness, sadness, surprise)
AffectNet [144]	W	Images collected from internet	450,000 subjects	Train. set: 23,901 images Val. 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)
DAiSEE [91]	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
UE-HRI [26, 27]	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
MHHRI [40]	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, neuroticism, openness, agreeableness, conscientiousness), 10-point Likert scale of engagement
Bhardwaj et al. ([30])		Online class	1000 participants	Emotion scores	10 EA	0-5 scale engagement level and emotions: angry, disgust, fear, sad, surprise and neutral

S - Spontaneous; **P** - Posed; **W** - in-the-Wild; **WE** - Web-based learning environment; **W** - in-the-Wild; **EA** - External annotator;

[249], JAFFE [135], CK+ [133], and RAF-DB [125] .

2.3.1 Engagement Measurement

Various approaches for measuring engagement include self-reports, experience sampling techniques, teacher ratings, interviews, and observations [79]. In addition, different indices (such as performance indices, number of clicks, and sensor data) have been used to assess engagement [234, 13, 235]. However, external observations ($n = 20; 43.48\%$), self-reported measures ($n = 9; 19.57\%$) and ratings are commonly used to measure engagement [223, 49] (Figure 2.4). Moreover, most publicly available engagement datasets were collected based on external observations by external annotators.

Self-reported measures are cheaper and easier to collect than external observations, which require more personnel to measure engagement [49]. Self-reports can be performed by self-annotating or completing questionnaires related to self-engagement [154]. However, self-reported measures are prone to Dunning-Kruger effects, as people are biased in recognizing self-competence [119, 160]. In addition, these measures are dependent strongly on participant compliance and diligence [71]. The bias associated with self-reported measures was also observed by Ramanarayanan et al. [170, 169].

Furthermore, observational measures limit the judgment quality of learners' actual effort, participation, or thinking [78, 162]. An external observer is an overhearer [186] 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 disengaged may be attempting to understand or relate new ideas to what they have learned [162]. 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 [78, 225].

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. [172] performed calculations on log data from courses in Moodle to evaluate engagement. Similarly, Hasnine et al. [92] calculated concentration indices, Apicella et al. [13] combined self-reported measures with performance indices, and Yue et al. [234] combined self-reported measures with quiz scores to assess engagement.

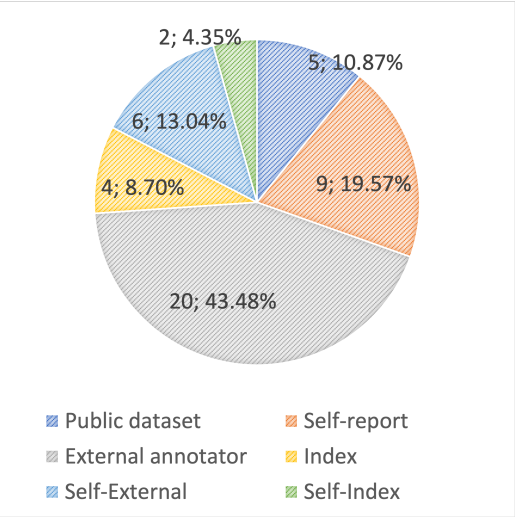


Figure 2.4: Pie chart of the engagement measurements and annotations used in the selected articles.

2.3.2 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 [48] or annotation tools such as CARMA [84], ANVIL [115], NOVA [24], and ELAN [226, 34] (see Appendix A.2.2).

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 [219, 11, 223, 240, 14, 199]. Cohen's kappa has also been used to evaluate the efficiency of classifiers for multiclass and imbalanced data [202].

The final label can also be determined by measuring intraclass correlations (ICCs) [85, 175] or by applying the majority-vote aggregation technique [237, 157, 240]. Highly consistent labeled data usually indicates high credibility [243].

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 [237]. Frame-based labeling is viewed as the easiest solution. However, this approach lacks continuous labels, which provide more precise information [197]. To address this issue, temporal

dynamics features need to be extracted [237].

However, in some cases, some frames are more significant for determining engagement levels, while other frames can mislead the final estimation result [247]. One solution for addressing this problem is applying an attention mechanism. The attention mechanism in deep learning directs attention to effectively choose important frames [213, 224].

Another labeling issue is a 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 [134]. Moreover, age can affect attention levels [134]. Therefore, collecting an engagement dataset representing learners' authentic internal states is challenging.

2.4 Machine Learning-based Methods for Automatic Engagement Estimation Module

RQ1.3: 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 [86]. 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 [189, 116, 245, 174], supporting the advancement of artificial intelligence in education applications [46, 156]. Therefore, automatic engagement estimation methods are generally 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 (Figure 2.5), 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 identified only using a nearly

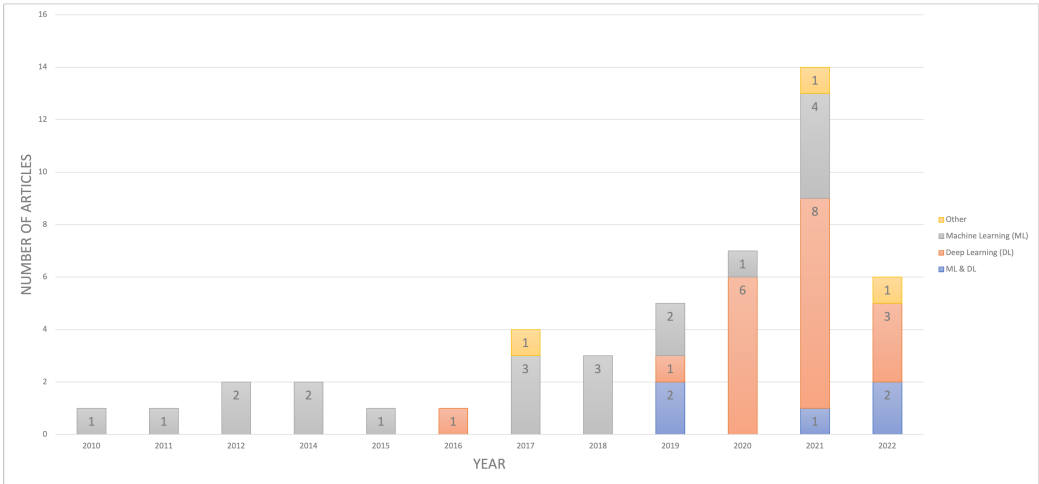


Figure 2.5: The general method used in the selected articles.

human-level understanding of data [86]. 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 [86]. For example, the input is presented as a visible layer to identify image features. 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 2.6 depicts a Venn diagram showing how deep learning is distinguished from classic machine learning.

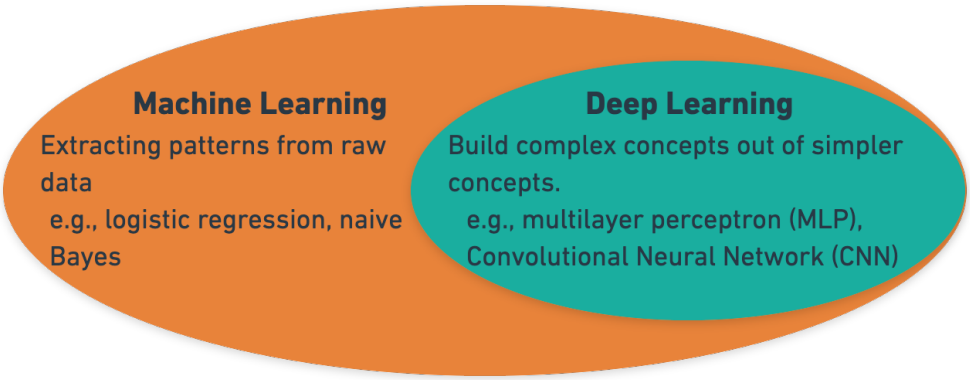


Figure 2.6: The difference between machine learning and deep learning.

Table 2.2: Face recognition tools for face detection and feature extraction

Tools name	Used in
OpenFace [22, 23]	[112, 175, 85, 136, 230, 241, 247, 204, 126, 73]
OpenCV	[232, 220, 54, 30, 92]
Dlib	[92, 141]
OpenPose [36]	[212, 246, 230, 247]
RetinaFace [59]	[197]
FasterRCNN [171]	[177]
faceAPI	[39]
Affectiva API in iMotion	[68]

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 [237]. Alternatively, statistical values can be extracted to obtain representation information from features in a given time window [94, 180, 237]. 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 [237].

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 [197, 223]. Additional low-level features include local binary patterns in three orthogonal planes (LBP-TOP), Gabor features, and box filters (BFs) [124].

High-level features are features extracted by aggregating low-level features [237], 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 [5, 17, 66, 216, 244]. These features can be extracted statistically or by using facial recognition tools, as shown in Table 2.2.

OpenFace is a popular computer vision toolkit for extracting facial features, including for automatic engagement estimation research (Table 2.2). OpenFace implements multitask cascaded convolutional networks (MTCNNs) [242] for face detection, constrained local models [21, 238] for landmark detection and tracking, eye rendering [228] for eye gaze estimation, and cross-

dataset learning and person-specific normalisation for facial action unit (FAU) detection. Figure 2.7 shows the output features extracted from OpenFace.

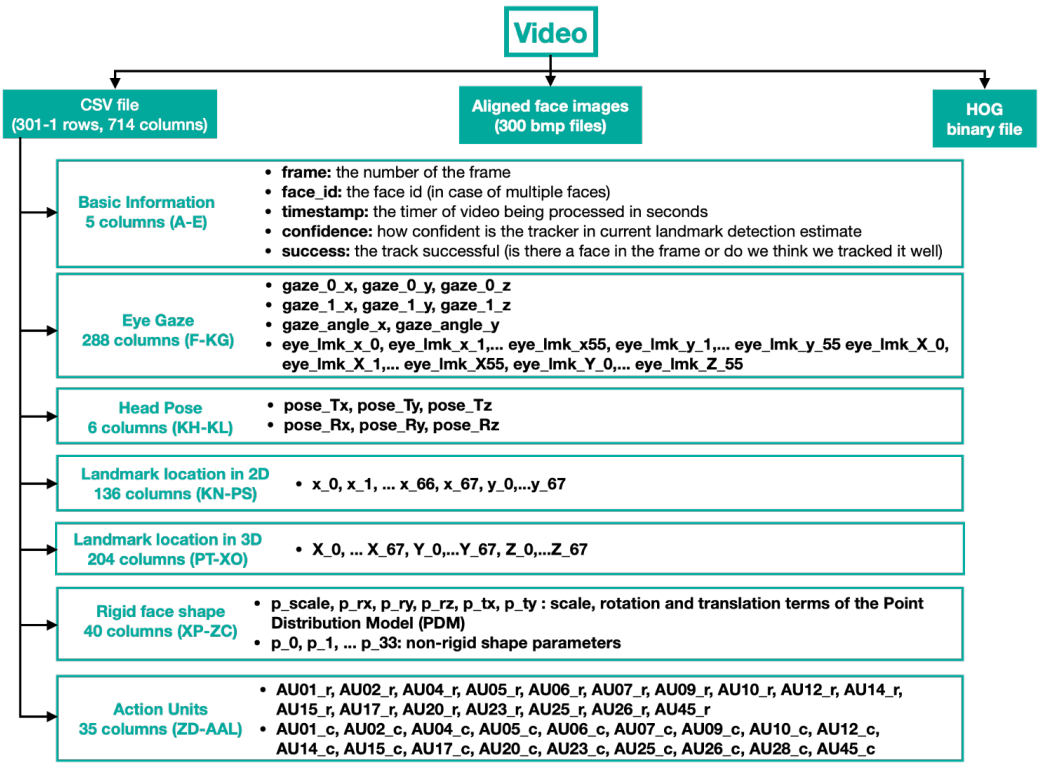


Figure 2.7: The output features of OpenFace [200].

In addition, the OpenCV[1] face detection library (Haar Cascade [215, 214, 185]) 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 [25]. Interested readers are referred to [218] 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. Data augmentation can be performed for image inputs by flipping (horizontally or vertically), cropping, scaling, or translating/rotating the images. As a result, the sam-

[1]https://opencv.org/

pling rate for the input can be increased by adding the augmented data to the original dataset [193, 199, 157].

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 [47], RELIEF-F [223], DeepLift [175], and recursive feature elimination random forests (RFE-RFs). Alternatively, ANOVA can be used to analyse the significance of labeled features [184].

Dimensional Reduction Dimensional reduction is the process of decreasing the dimension of the input feature to prevent overfitting [237]. 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) [197, 219] and forward feature selection (FFS) [11]. However, the 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 [237].

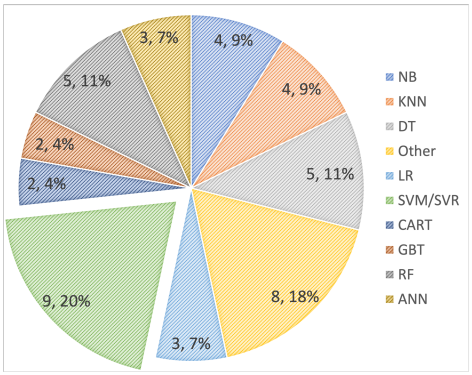
Addressing Imbalanced Data One major issue with engagement datasets is imbalanced data that are severely skewed towards the majority class [237]. Imbalanced class labels often occur because disengagement is rarely observed in continuous labeling. Many methods have been proposed to address this issue [82, 45, 83, 67]. There are three categories of re-sampling techniques [25]: 1) under-sampling 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) [45]; and 3) hybrid methods that combine both sampling methods [83, 45]. Moreover, continuous scales may be discretized into groups [178, 176], and weighting techniques [67, 129] have also been used to address this problem.

Classic Machine Learning Methods

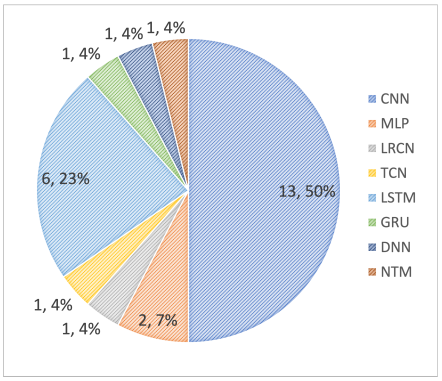
Engagement is estimated by calculating probabilities. To calculate the engagement probability, several classic machine learning methods can be utilized, such as the support vector machine (SVM) and its variations (including support vector regression (SVR)), naive Bayes (NB), decision trees (DTs), logistic 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) [97, 101] (as used in [50, 146, 172]), the Computer Expression Recognition Toolbox (CERT) [130] (as used in [223]), and the MATLAB library ([44, 11].

Between 2010 and 2022, classic machine learning methods dominated the automatic engagement estimation literature (Figure 2.5), especially SVMs (2.8a). Note that some of the selected articles examined more than one algorithm. Therefore, the totals in Figure 2.8a do not correspond to the number of selected articles (see Appendix Table 5).



(a) The use of classic machine learning methods.



(b) The use of deep learning methods.

Figure 2.8: Pie chart of the use of a) classic machine learning and b) deep learning methods for automatic engagement estimation.

Deep Learning Methods

This section briefly introduces some deep learning methods, including those used in the selected articles. For a more detailed explanation of deep learning techniques (especially for face recognition), interested readers are referred to [218, 80, 124].

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 formed by combining many simpler functions to map some input values to output values [86]. 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 layers are called the input and output layers, respectively, while the number of hidden layers

may vary, determining the model's **depth**. Furthermore, depending on the number of inputs and outputs, each layer may contain more than one unit. This algorithm has also been used in automatic engagement estimation for performance comparison with other algorithms [27, 197, 175].

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 [86]. 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 [86].

CNNs are currently one of the most popular methods in different fields (Figure 2.8b). This technique has been widely used in various computer vision applications, including image classification [93], semantic segmentation [152], object detection [198], face recognition [159], spatiotemporal feature learning [208, 99, 106, 237, 177, 2, 15, 234], and automatic engagement estimation (see Appendix Table 5).

CNNs are popular because they can be highly modified and pre-trained. Some CNNs include AlexNet [117], i3D [38], VGG16 [194], and ResNet [93, 198].

The inputs to a CNN are usually greyscaled or RGB images. Using 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 [143, 220]. 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 [202].

Recurrent Neural Network (RNN) A facial expression changes through three stages, i.e., onset, apex, and offset [131]. 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 [108]. RNNs capture information at earlier and later time steps by remembering each piece of information over time [1]. Therefore, this algorithm has become a more popular automatic engagement estimation method (see Appendix Table 5).

Some types of RNNs include long short-term memory (LSTM) [96] [234, 27, 56, 127, 197, 73], gated recurrent units (GRUs) [27], and network Turing machines (NTMs)[168] [136].

However, despite advantages such as considerable computational power in

temporal processing models and applications, RNNs are difficult to train in practice due to network instability [1]. 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 [1].

Therefore, the attention mechanism was introduced to learn to associate the elements in the sequence C with the elements in the output sequence [18]. The attention mechanism essentially determines a weighted average that is used to focus on specific parts of the input sequence at each time step [86]. Although the attention mechanism was originally introduced in the context of machine translation [18], it has also been utilized in DL applications for automatic engagement estimation [127, 197, 141, 193].

Other Classifiers Other neural network techniques that have been used for automatic engagement estimation include the fuzzy min-max neural network (FMMNN) classifier [195, 81], which was implemented by [235] for automatic engagement estimation, the deep belief network [95], which was used in [60], and linear discriminant analysis (LDA) [13, 219].

Fine-Tuning and Transfer Learning Techniques

One fine-tuning technique for addressing insufficient training data is applying transfer learning, which utilizes networks pre-trained on a large number of images [28, 218]. Various models have been trained on large face image datasets. For example, [197] used AffectNet [144] and 300W-LP [249], which were trained on ResNet50, for transfer learning. The pre-trained models help the engagement estimation network learn general features related to face identification [237]. As mentioned in Section 2.3, other large datasets that have been used for transfer learning include FER-2013 [87], VGGFace [159], VGGFace2 [35], FaceNet [188], AffectNet [144], 300W-LP and AFLW2000 [249].

Performance Metrics

To judge the automatic engagement estimation performance, the prediction results should be compared with human judgments in the dataset [223, 237]. In the machine learning pipeline, performance metrics are used to monitor and measure the performance of a model depending on the task. Automatic engagement estimation problems can be seen either as classification or regression tasks. An engagement estimation is a classification task if the engagement is estimated in discreet classes, e.g., low engagement class vs

high engagement class. Otherwise, an engagement estimation is a regression task when continuous output is desired. Some metrics used to measure the performance of regression tasks are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which mainly calculate 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, a confusion matrix is also used to visualize the ground-truth labels versus the predicted results in a table.

The 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, the Precision/Recall (PR) trade-off curve (used in [123]) and F1-score [184] 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, the 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 [94, 123] and 2-alternative forced choice (2AFC) [223].

AUC-ROC visualizes the classification performance based on correct and incorrect classifications (Figure 2.9). 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 [33]. The AUC scores between $0.7 - 0.8$, $0.8 - 0.9$, and > 0.9 are considered acceptable, excellent, and outstanding, respectively [137, 126].

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall/TP \text{ rate} = \frac{TP}{TP + FN}$$

$$FP \text{ rate} = 1 - TP \text{ rate} = \frac{FP}{TN + FP}$$

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall}$$

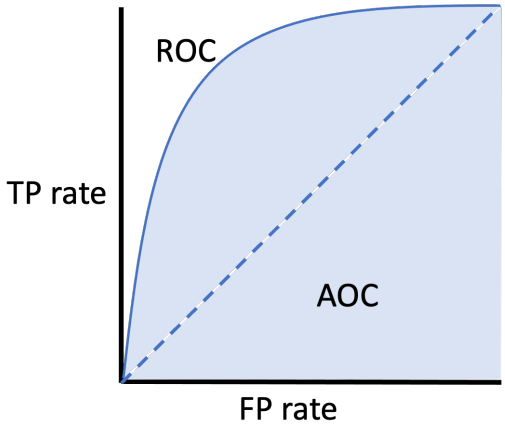


Figure 2.9: AUC-ROC curve illustration.

The 2-alternative forced choice (2AFC) [138, 206] 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) [206] and specificity and sensitivity [237] are also used in the engagement estimation literature (see Appendix Table 5).

2.5 Chapter Conclusion

This chapter 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 [158] to answer three sub-RQs of RQ1:

- RQ1.1: How should the type of engagement to be measured be defined?
- RQ1.2: What datasets are suitable for developing automatic engagement estimation methods?

- RQ1.3: What automatic engagement estimation methods have been developed in the literature?

The results and discussion in Section 2.2, 2.3, and 2.4 with the presented information, figures, and tables aims at providing new researchers or educators with insight on automatic engagement estimation to enhance smart learning with automatic engagement recognition methods.

To answer RQ1.1, we examined the definitions of engagement used in the selected articles and introduced an engagement definition taxonomy (Figure 2.2) 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, physiological, 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 defining what type of engagement is being measured depends on engagement cues used, what stimulus was presented to the participant during data collection, and what physical or cognitive behaviours were observed. We believe 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 RQ1.2. 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 led 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 RQ1.3, we discuss machine learning-based methods that 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.

2.5.1 Remaining issue

The combination of a clear definition of engagement 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 educational settings remains the challenge, especially in addressing the main research question, "How do educators or education institutions apply automatic engagement estimation in their distance learning process?" Therefore, experiments on a deep learning-based automatic engagement estimation module will be discussed in the subsequent section as the further step to address the main RQ. Furthermore, the ethical impact remains unaddressed in the existing works. We believe that the haphazard implementation of this technology could abuse user privacy and ethics.

Chapter 3

Automatic Engagement Estimation Model

This chapter is an updated and abridged version of the following publications:

1. S. N. Karimah and S. Hasegawa, "A Real-time Engagement Assessment in Online Learning Process Using Convolutional Neural Network," in The 12th Asian Conference on Education (ACE2020) , Jan. 2020, pp. 437–448, doi: 10.22492/issn.2186-5892.2021.39,

2. S. N. Karimah, T. Unoki, and S. Hasegawa, "Implementation of Long Short-Term Memory (LSTM) Models for Engagement Estimation in Online Learning," in 2021 IEEE International Conference on Engineering, Technology & Education (TALE), Dec. 2021, pp. 283–289, doi: 10.1109/TALE52509.2021.9678909.

3.1 Chapter Introduction

According to the systematic review results in Section 2, computer vision-based methods are the most popular methods in the literature for automatic engagement estimation. Therefore, in this thesis, we follow the majority of studies that use visual-based analysis from video to estimate emotional engagement.

Computer vision-based methods offer several ways to estimate learners' engagement by optimizing the appearance features such as body pose, eye gaze, and facial expression. Grafsgaard et al. [88], Whitehill et al. [223], and Monkaresi et al. [146] using machine learning to estimate engagement from facial expression features. They used machine learning toolboxes, e.g., Computer Expression Recognition Toolbox (CERT) [130] and WEKA [97,

101], to track the face and classification. However, using the toolboxes for engagement estimation will automate a part of the classification process but not the implementation in the real-time education process since humans manually input the extracted features. On the other hand, Nezami et al. [148, 143] and Dewan et al. [60] use deep learning to build their own classification model to estimate the engagement of online learners which possibly enables to make the pre-processing both in the implementation process and the training process is done in the same way so that the input for engagement prediction is in the same distribution as the input for classification model training. Thus, the deep learning methods have been the state-of-the-art (SOTA) for automatic engagement estimation in the past five years (Figure 2.5).

Following SOTA, in this chapter, we build engagement classification models using deep learning to classify the real-time image into very engaged, normally engaged, or not engaged classes. Furthermore, the existing studies on automatically recognising learner engagement have focused on the accuracy performance of machine learning-based recognition to estimate learner engagement. However, a development framework is required to be able to use the advantage of automatic engagement estimation in actual distance learning settings. Therefore, this chapter proposes a framework for the practical use of a real-time engagement estimation to assess the learner's engagement state while participating in a distance learning process. The framework depicts the end-to-end process of an engagement estimation tool in an online learning management system (LMS) or a web-based environment, where the input is the real-time images of the learners from a webcam.

Firstly, we conducted sequence-based experiments considering that engagement as a dynamic inner state. Four long short-term memory (LSTM) models were investigated by experimenting with six pre-processing scenarios. The experiment result shows the best combination or order of pre-processing methods with the models. The models experimented on Dataset for Affective States in E-Environments (DAiSEE) [91] with modification. As mentioned in Section 2.3, DAiSEE is the most popular publicly available dataset used in learner engagement studies.

Secondly, a face recognition and engagement classification model to analyse learners' facial features was developed, leading to the development of an automatic engagement estimation module. Using the same public engagement dataset, a convolutional neural network (CNN) was adopted to classify them into one of the three engagement classes: very engaged, normally engaged, or not engaged. At the end of the chapter, we propose an automatic engagement estimation framework for real-time implementation and discuss the estimation performance and practicality, leading to the following chapter.

3.2 Dataset Overview

The Dataset for Affective States in E-Environment (DAiSEE) is a publicly available dataset with a multi-label video classification of 112 participants [91]. The data were collected in an "in-the-wild" environment to simulate the real-world distance learning environment where a learner may join the online class anywhere. To limit the occurrence of the Hawthorne effect [145, 139], the participants were recorded without being trained, and no parameters for the experiment were set. Therefore, the illumination of the videos in the dataset varies in three different settings, i.e., light, dark, and neutral.

Each participant watched one educational video and one recreational video to capture both focused and relaxed settings. The total length of the two videos was 20 minutes. The recorded videos of each participant (approximately 13 to 20 minutes) were split into 10-second video snippets resulting in 9068 video snippets in the dataset. 8925 snippets were annotated based on the "wisdom of the crowd" of 10 annotators. To obtain the ground truth label of each video snippet, the Dawid-Skene [53] vote aggregation algorithm was used.

Image features and high-level features were extracted using our self-built Python script and OpenFace 2.0 [23], respectively. Figure 3.1 shows the structure of DAiSEE and two feature extraction methods we employed.

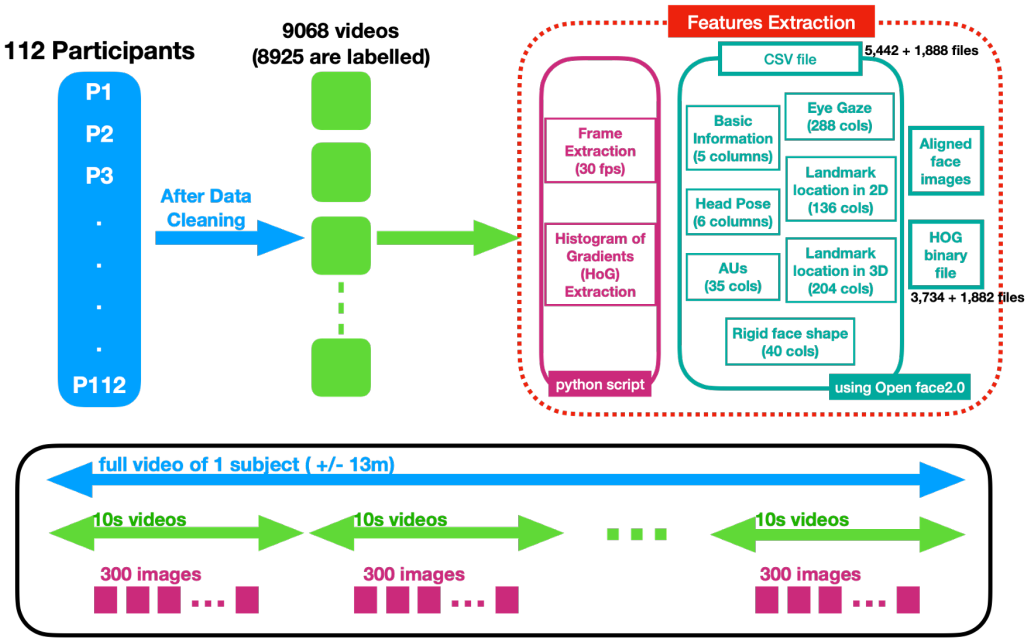


Figure 3.1: The structure of DAiSEE.

The dataset was divided into training ($N = 5358$), validation ($N = 1429$), and test sets ($N = 1784$), as shown in Table 3.1. Note that the engagement label distribution, as shown in Table 3.1, is based on the amount of actual data we obtained after feature extraction.

Table 3.1: Engagement Label Distribution of the dataset

Label	0	1	2	3	Σ
Training	34	213	2617	2494	5358
Validation	23	143	813	450	1429
Test	4	84	882	814	1784
Σ	61	440	4312	3758	8571

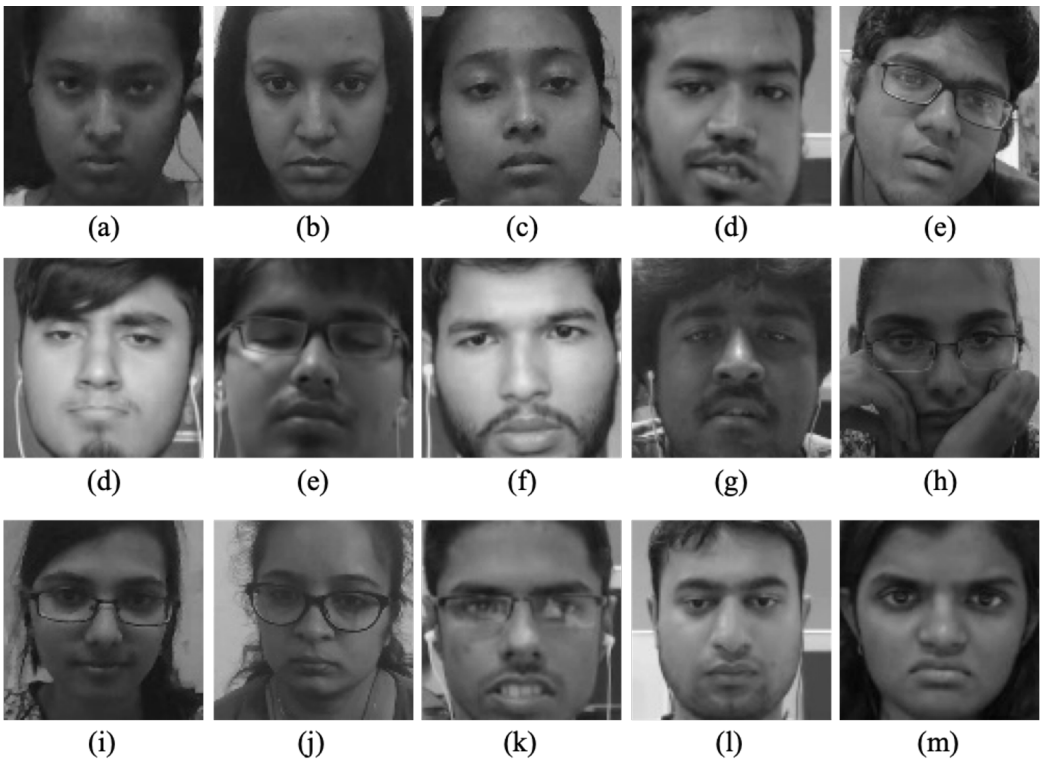
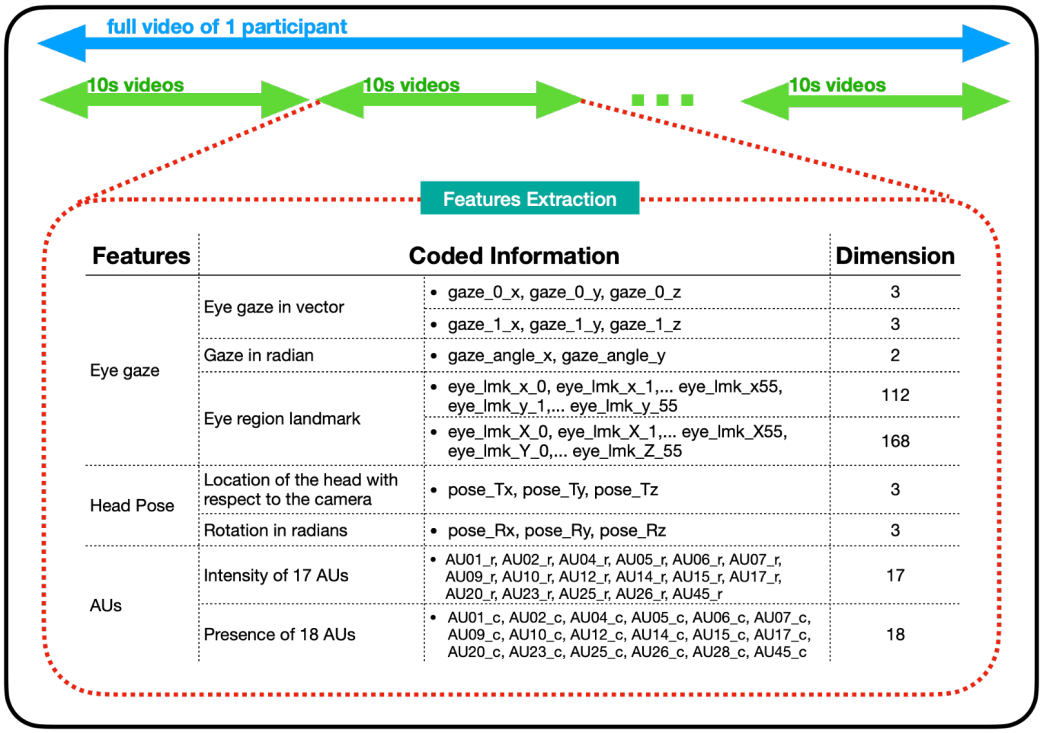


Figure 3.2: Samples of extracted images from DAiSEE. (a)-(e), (d)-(h), and (i)-(m) are images with labelled as *very-engaged*, *not-engaged* and *normal-engaged*, respectively.

We modified the label by combining the 0 and 1 labels to have three-level engagement states, which represent *Not Engaged, Normally Engaged*, and *Very Engaged*. The rationale for this modification is that because engagement

is a subtle state, annotators have different intuitions when discriminating between very low and low engagement levels, and the two labels are visually interchangeable. Another reason for this label modification was the severe imbalance distribution, as the number of *very low* and *low* level classes was far lower than the high and very high-level classes. The three-level label modification has also been done in [60]. The sample images extracted from the Python script are shown in Figure 3.2, whereas Figure 3.3 shows the extracted features and information extracted from OpenFace.



Features	Coded Information		Dimension
Eye gaze	Eye gaze in vector	• gaze_0_x, gaze_0_y, gaze_0_z	3
		• gaze_1_x, gaze_1_y, gaze_1_z	3
	Gaze in radian	• gaze_angle_x, gaze_angle_y	2
	Eye region landmark	• eye_lmk_x_0, eye_lmk_x_1,... eye_lmk_x55, eye_lmk_y_1,... eye_lmk_y_55	112
		• eye_lmk_X_0, eye_lmk_X_1,... eye_lmk_X55, eye_lmk_Y_0,... eye_lmk_Z_55	168
Head Pose	Location of the head with respect to the camera	• pose_Tx, pose_Ty, pose_Tz	3
	Rotation in radians	• pose_Rx, pose_Ry, pose_Rz	3
AUs	Intensity of 17 AUs	• AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU07_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU23_r, AU25_r, AU26_r, AU45_r	17
	Presence of 18 AUs	• AU01_c, AU02_c, AU04_c, AU05_c, AU06_c, AU07_c, AU09_c, AU10_c, AU12_c, AU14_c, AU15_c, AU17_c, AU20_c, AU23_c, AU25_c, AU26_c, AU28_c, AU45_c	18

Figure 3.3: Extracted features with respect to the scheme of one participant in DAiSEE.

3.3 Implementation of Long Short-Term Memory (LSTM) Models for Engagement Estimation

Since engagement is a dynamic state that frequently changes and is influenced by a sequence of previous states, we employ LSTM models, which enable us to consider a time-series scenario. This chapter investigates the suitability of

four Long-Short-Term Memory LSTM models: single LSTM, stacked LSTM, Bidirectional LSTM (Bi-LSTM), and Bi- LSTM with additional precedent neural network layers (Multilayer Bi-LSTM) for engagement estimation. The practical contribution of this chapter is two-fold: (1) provide baseline results of time-series-based engagement estimation on the Dataset for the Affective States in E-Environments (DAiSEE) following the result of previous researches [91, 98, 127, 246]; (2) suggest pre-processing combinations method, i.e., downsampling, oversampling, and Principal Component Analysis (PCA), to improve the prediction accuracy.

3.3.1 Dataset modification

We modified the set division to ensure that the training set and the validation set were treated in the same way in pre-processing. Instead of using separate training and validation sets, we first combined the two sets to create a training set. The validation set was 20% of the new training set obtained automatically during model compiling and training. The modified label and set distribution are shown in Figure 3.4, in which the detail $N = 2944(43.377\%)$, 3430 (50.538%), 413(6.085%) for *Very Engaged, Normally Engaged*, and *Not Engaged* labels in the training set (total 6787), and $N = 814(45.628\%)$, 882(49.439%), 88(4.933%) in the test set (total 1784).

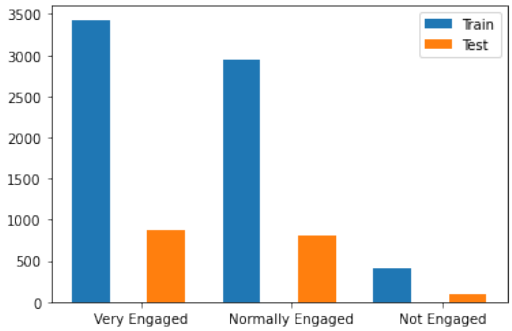


Figure 3.4: Modified engagement label distribution for LSTM models experiment.

3.3.2 Feature extraction

Since the release of OpenFace [22] and OpenPose [36] in 2017, the trend in engagement estimation research has been to use high-level features such as face analysis results rather than low-level features such as pixel-based features [110]. OpenFace and OpenPose are open-source toolkits used for facial and

body behavior analysis. OpenFace is capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation, while OpenPose can estimate 2D body pose (see Section 2.4).

Following [42, 150, 112], we used OpenFace 2.0 [23] toolkit to extract the information of head pose, eye gaze, and gaze direction [228], and Action Units (AUs) [20] for automatic engagement estimation. We configured OpenFace 2.0 and used Multi-task Cascaded Convolutional Network (MTCNN) as a face detector and Convolutional Experts Constrained Local Model (CE-CLM) as a landmark detector. They obtained 329 dimensions of features for each video snippet (Figure 3.3). These features include:

1. Eye gaze contains two 3D eye gaze direction vectors, one 2D eye gaze direction in radians, and fifty-six 2D and 3D eye region landmarks (total 288 features).

2. Head Pose contains six dimensions in total, with the detailed 3D location of the head with respect to the camera and 3D rotation in radians.

3. Facial Action Units (AUs) contain the intensity (from 0-5) of 17 AUs and the presence (0 for the absent and 1 for the present) of 18 AUs (total 35 features).

3.3.3 Pre-process

We aim to use the averaged intuition of a video to express the learning situation and investigate the effect of the pre-processed method on the model. Since the video rate is 30 frames per second (FPS), there are 300 frames extracted in 10 seconds. We averaged all of the frames in a video, used the averaged value, and considered the resulting frame as one time-step in sequence data. We then added the label to the averaged frame and concatenated it with the averaged frames of all videos into a file. The rationale for the averaging is that some video snippets were not exactly 10 seconds long.

We used both undersampling and oversampling techniques in the training data since the number of data belonging to *Not Engaged* class is far fewer than the other two classes (Figure 3.4) (Figure 3.5). First, we undersample the Very Engaged and Normally Engaged classes by 50%, then oversample the Not Engaged class using Synthetic Minority Oversampling Technique (SMOTE)[45].

For data normalization, we used min-max normalization to re-scale the range of features to range in $[0, 1]$, where the general formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{3.1}$$

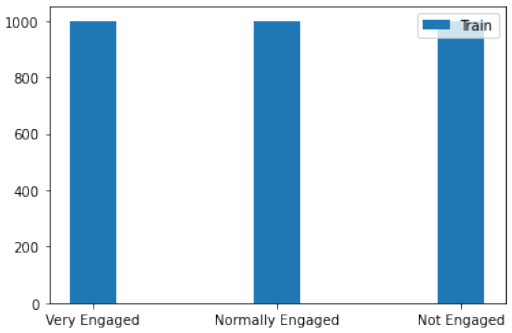


Figure 3.5: Resampled distribution.

where x and x' are the original and the normalized values, respectively. To reduce the dimension of the dataset while preserving as much information as possible, we then used Principle Component Analysis (PCA) by $n = 250$.

We experimented with the dataset pre-processing in six scenarios to investigate which combination or order of pre-processing methods yielded the best prediction accuracy. All the scenarios were applied with re-sampling, except for scenario 6.

- **Scenario 1** : no normalization, no PCA.
- **Scenario 2** : only apply normalization.
- **Scenario 3** : only apply PCA with $n = 250$.
- **Scenario 4** : apply PCA first, then Normalization.
- **Scenario 5** : apply normalization first, then PCA.
- **Scenario 6** : apply scenario 5 with the original data distribution (no undersampling nor oversampling).

The summary of the data pre-processing is shown in Figure 3.6.

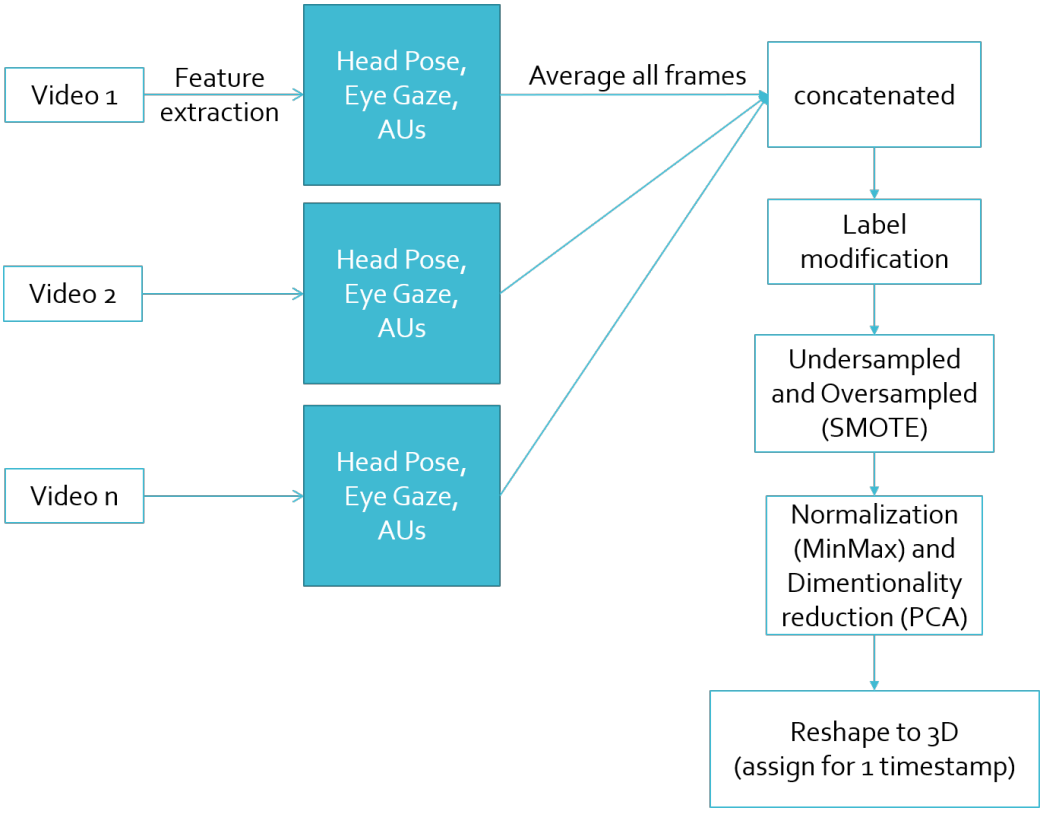


Figure 3.6: Data pre-process of Scenario 5.

Table 3.2: Experiment Results

Scenario	Single-LSTM		Stacked-LSTM		Bi-LSTM		Multilayer Bi-LSTM	
	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE
	Validation/Training							
1	0.000/0.423	0.349/0.208	0.000/0.411	0.347/0.208	0.000/0.417	0.667/0.389	**0.428/0.545**	**0.239/0.185**
2	0.567/0.639	0.193/0.158	**0.627/0.684**	**0.174/0.142**	0.498/0.601	0.214/0.170	0.518/0.709	0.208/0.134
3	0.518/0.692	0.213/0.140	**0.627/0.782**	**0.196/0.102**	0.032/0.538	0.641/0.305	0.576/0.639	0.190/0.156
4	0.611/0.643	0.167/0.157	0.631/0.653	0.177/0.152	0.536/0.641	0.202/0.158	**0.902/0.944**	**0.054/0.030**
5	0.732/0.778	0.132/0.110	**0.800/0.914**	**0.110/0.050**	0.786/0.861	0.110/0.077	0.729/0.764	0.132/0.111
6	0.454/0.753	0.245/0.118	**0.507/0.873**	**0.284/0.065**	0.501/0.827	0.259/0.088	0.464/0.753	0.260/0.114

3.3.4 Experiment Setup and Result

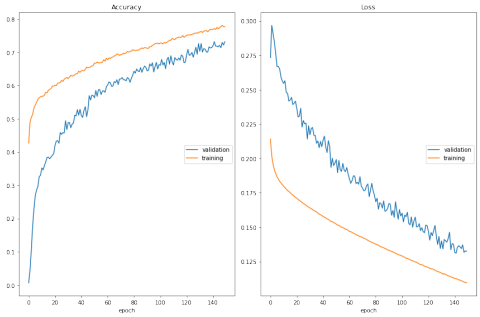
We carried out experiments on the modified dataset to evaluate the four models. The experiments were conducted using the following hardware and

software settings:

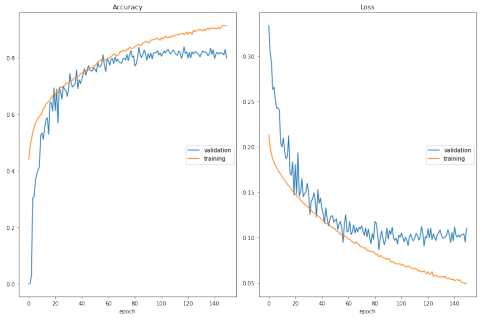
1. Processor Intel Core™ i9-10900K CPU 3.70 GHz.
2. 32.0 GB RAM.
3. Windows 10 Pro 64-bit, version 20H2.
4. Keras 2.3 using Tensorflow 2.1 backend, and Python 3.7.
5. NVIDIA GPU Computing Toolkit with CUDA version 10.2.

All the experiments were compiled using the mean squared error loss function and Adam optimizer with $1e-3$ learning rate and trained in 150 epochs. Table 3.2 shows the results of all experiment scenarios; the best performance of each model is shown in Figure 3.7.

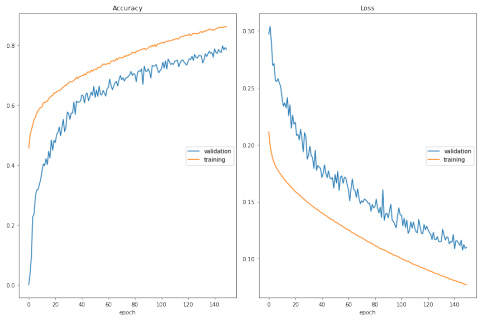
Overall, the Multilayer Bi-LSTM in the scenario 4 yield shows the best performance, with a validation accuracy of 0.902. The Stacked-LSTM is the most robust compared to other models because it shows the best performance in scenarios 2, 3, 5, and 6. Scenario 5 works well on all models, with a validation accuracy of 0.732, 0.800, 0.786, and 0.729 for Single-LSTM, Stacked-LSTM, Bi-LSTM, and Multilayer Bi-LSTM, respectively.



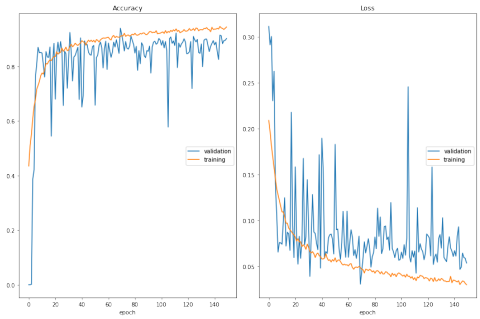
(a) Single-LSTM Scenario 5.



(b) Stacked-LSTM Scenario 5.



(c) Bi-LSTM Scenario 5.



(d) Multilayer Bi-LSTM Scenario 4.

Figure 3.7: Plot Accuracy and Loss (MSE) of the best performance of each model.

3.4 CNN Classification Model for a real-time Engagement Estimation Tool

This work employed a convolutional neural network (CNN) for engagement classification using the image features obtained from Viola-Jones (V&J) face detection. We use CNN because it is relatively simple and one of the deep learning methods broadly used in literature [90, 124, 147, 148]. Furthermore, we believe simplicity and cost efficiency are the keys to a reliable implementation of engagement estimation in the real-world online learning process.

3.4.1 Pre-process

The pre-processing comprises V&J face detector [215], where rectangle features are used to detect the presence of that feature in the given face images. Figure 3.8 shows three types of rectangle features used in V&J face detection, i.e., two-rectangle feature, three-rectangle feature, and four-rectangle feature. The sum of pixels under the white rectangle is subtracted from the sum of pixels under the black rectangle, resulting in a single value in each feature.

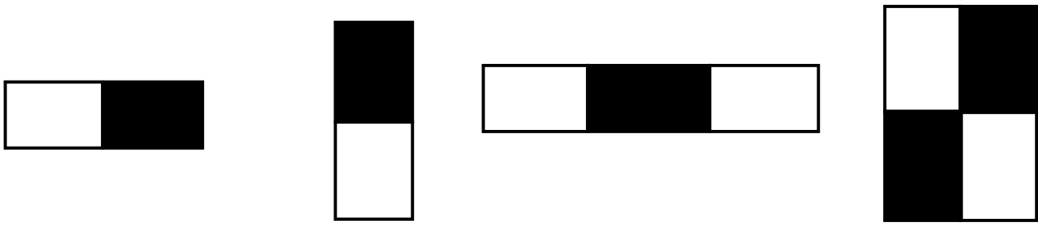


Figure 3.8: Rectangle features used in V&J face detection.

The rectangle features are computed rapidly using integral images to be processed in real-time [214, 215]. Given that the base window is 24x24, the dimensionality of the set of rectangle features is quite large, e.g., 160,000+ features. Therefore, Adaboost is used for dimensionality reduction (from 160,000+ features to 6,000 features) and to find the single rectangular feature and threshold that best separates the positive (faces) and negative (non-faces) images. Then, all the features are grouped into several stages using a cascade classifier. Each stage has a certain number of features to form complete face images while discarding the negative images. The face images are then represented in a rectangular region of interest (RoI) to be then fed to the Neural Network for training.

The classification used a typical CNN architecture which contains an input layer, multiple hidden layers, and an output layer. The hidden layers combine convolutional layers, activation layers, pooling layers, normalization layers, and fully connected layers that we classified into convolution blocks and fully connected blocks as depicted in Figure 3.9.

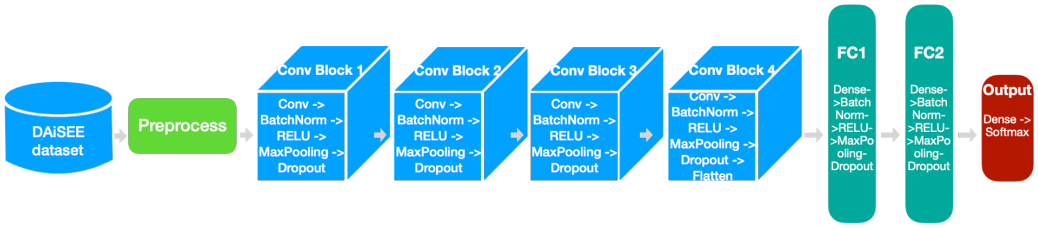


Figure 3.9: CNN architecture used in this work.

3.4.2 Experiment Setup and Result

To build an engagement estimation tool prototype, we experimented with 5045 face images for training and 1525 face images for validation tests. We then exported the model and the weights into JSON and H5 files, respectively. We set the number of convolution and filter layers in the convolutional blocks as the primal hyper-parameters, i.e., 64 (3,3), 128 (5,5), 512 (3,3) 512 (3,3) for convolutional blocks 1,2,3, and 4, respectively. We use Dense layers 256, 512, and 3 for fully connected blocks and softmax layers. Other hyper-parameters we also set are Max Pooling (2,2), dropout (0.25), and rectified linear unit (RELU) activation in all convolutional blocks, while for optimisation, we used Adam optimizer with learning rate 0.0005 and L2 regularization 0.0001. From the network and hyper-parameters set above, the training accuracy is 59.25%, and the validation accuracy is 56.91%.

3.4.3 A Real-time Engagement Assessment Framework

There appears to be a practical knowledge gap in the prior research (Section 2). Most automatic engagement estimation studies have primarily focused on the theoretical aspects of the machine learning-based method, as we have done in the previous sections of this chapter. However, there are very few practical studies or action research in the field of education, especially distance learning.

For example, when traditional face-to-face methods are transferred to distance education, we assume that the engagement recognition should be done in real-time, as in the traditional classroom. Therefore, to support the

real-time implementation in an actual distance learning setting, we propose a framework as shown in Figure 3.10. The general ideas of the proposed framework are: 1) a web-based application to enable multi-platform access, and 2) using the same face detection and feature extraction methods in both training the model and real-time implementation. Our proposed framework allows the generated log file can be downloaded anytime by the educator to evaluate their teaching or course planning.

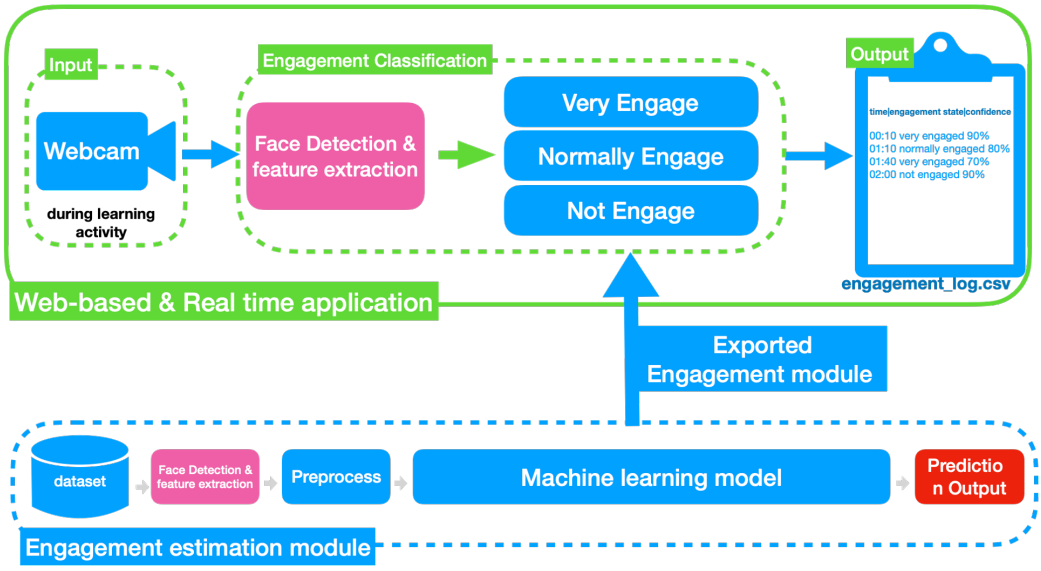


Figure 3.10: Proposed framework for a real-time engagement assessment.

The trained model and weight were exported and served in Flask Python to build the web application using the CNN experiment result. The screenshot of the running application is shown in Figure 3.11.

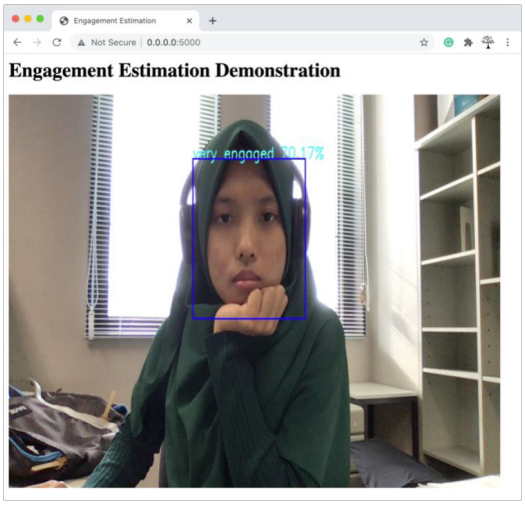


Figure 3.11: Screenshot of the running engagement estimation tool with CNN classification module.

3.5 Performance Evaluation for Practical Implementation

In this section, we evaluate the performance of LSTM and CNN models to build an engagement estimation model by comparing the prediction accuracy and the runtime on the test set of the DAiSEE and the runtime. In addition, we also compare it with a classic machine learning method. Since we want to solve the multi-class classification task, choose Logistic Regression (LR) among the classic machine learning methods from the Scikitlearn library.

For comparison purposes, we modify the input features and simplify the model to evaluate the test set. The experiment is conducted using the same hardware and software as the previous LSTM experiment (Section 3.3.4) in scenario 5 (normalisation first, then PCA). The input is an average of all frames of data (1 timestep). The training settings are in single validation (SV) with Training/Validation/Test set (5467/1703/1782) and 5-Fold stratified cross-validation (8952), where the results are shown in Tables **??**.

Table 3.3: Performance Evaluation: Single Validation (SV).

Models	Train. acc.	Val. acc.	Test acc.	Training time (mm:ss)
LR-SV	0.522	0.522	0.329	**00:00.2**
CNN-SV	0.919	0.868	0.3424	02:10.1
SingleLSTM-SV	0.944	0.858	0.3779	03:02.3
StackedLSTM-SV	0.935	0.92	0.3484	01:19.9
BiLSTM-SV	0.969	0.885	0.3567	03:01.9
MultilayerLSTM-SV	0.843	0.853	0.3499	03:01.6

Table 3.4: Performance Evaluation: Single Validation (CV).

Models	Mean acc.	Training time (mm:ss)
LR-CV	0.502	00:01.7
CNN-CV	0.822	09:09.3
SingleLSTM-CV	0.82	05:21.1
StackedLSTM-CV	0.82	05:21.1
BiLSTM-SV	0.811	05:31.5
MultilayerLSTM-CV	0.747	06:56.8

3.6 Discussions

We have presented the efficacy of the four LSTM models in the DAiSEE dataset to estimate the engagement state of learners, extending the existing research, such as has been done in [42]. In this work, we use the DAiSEE dataset because it includes an engagement label and has been used for engagement estimation research in literature (Section 2.3). However, we found that dataset preparation to build the classification model is the most challenging issue. For example, there is a significant difference in the number of images between the classes. As shown in Table 3.1, the number of images in intensity 3 (i.e., very engaged class) is much larger than in other intensities/classes, especially intensity 0 and 1 (i.e., not engaged class). Furthermore, as shown in Figure 3.2, it is difficult to distinguish between the images with different class labels. Additionally, the different illumination settings in the data cause the extracted features, not to be in the same distribution. Therefore, we introduce pre-processing scenarios in Section 3.3.3 to investigate the different pre-processing methods affecting the engagement data behaviour during the training.

The best performance (0.902 accuracy) was achieved using Multilayer-LSTM with scenario 4, where the re-sampled data was pre-processed with PCA before applying the normalization. The other three LSTM models also performed well in scenario 5, with validation accuracy of 0.732, 0.800, 0.786, and 0.729 for Single-LSTM, Stacked-LSTM, Bi-LSTM, and Multilayer Bi-LSTM, respectively.

One of the hypotheses for these good results is to consider one time step by averaging all the frames in a video. During a labeling process, the annotators are unable to recognize detail behaviors, rather annotate based on the overall impression which may be shown in average data. However, the average of all the frames may have resulted in data loss or merged the important and unimportant information resulting in the failure of the model to capture the general information [150].

Future work should consider more important time steps, which may possibly give more representation of the subtle change in time and thus potentially avoid such problems. In addition, future work may put more attention on the pre-processing to the occurrence where the face is widely turned away such that OpenFace might not be able to recognize the face feature. Such extreme behaviour may result in high gaps in data and bias in the training process.

When the aforementioned issue is addressed, we believe that LSTM provides a promising solution for more accurate estimation. However, when it comes to real implementation, LSTM is computationally expensive and unstable, especially for real-time implementation, as shown in the performance evaluation in Table **??**. Compared to CNN and Logistic Regression, LSTM takes a longer runtime to predict the test set.

Despite the low accuracy of the evaluation performance, we still can see that CNN is a more practical deep learning network than LSTM. However, classic machine learning (Logistic regression) shows the best practice in terms of runtime.

3.7 Chapter Conclusion

In this chapter, we build deep learning-based automatic engagement estimations. LSTM and CNN were experimented with to build the classification models trained on DAiSEE. Considering the sequential characteristic of the engagement state, we apply LSTM models on the DAiSEE and suggest the data pre-processing scenarios in Section 3.3.3. The models were trained on the DAiSEE in six pre-processing scenarios (Section 3.3.3) to determine which combination or order of pre-processing methods work best with the

models. The results show that scenario 5 works well, performing > 0.729 in all models and that Multilayer Bi-LSTM achieved the best performance in scenario 4 (0.902 accuracy). This work is the first stage for reliable automatic engagement estimation in online learning, taking into account the sequential characteristic of the engagement state. Further improvement, as discussed in Section 3.6, needs to be done for more robust estimation and applicability in the real-world distance learning system.

Meanwhile, although the sequential characteristic of engagement is missing, the CNN model is more practical for tool development. Therefore, this chapter uses the CNN classification model to bridge the practical knowledge gap.

A framework of automatic engagement assessment of learners in online learning has been introduced to give an image for implementing real-time engagement assessment in an actual distance learning scenario (Figure 3.10). The term automatic not only regards the classification method but also includes the end-to-end real-time process when the learner is conducting online learning. The face detection methods in the model development and online application should be the same to yield the input for the engagement classification in the same distribution as the input for building the classification model.

The implementation of the framework is the engagement estimation tool prototype as a web-based application, as shown in Figure 3.11. The proposed framework shows that the pre-process to build the classification model from the dataset, and the online implementation of learners' engagement estimation must be done in the same way. Therefore, our experiment on LSTM is not suitable for real-time applications since it uses OpenFace for feature extraction, which cannot be run simultaneously when the system runs online. In contrast, the CNN model allows real-time estimation since the input to be estimated in the classifier is an image, which can be done using OpenCV. Therefore, the CNN model is used to evaluate the proposed framework suitably for real-time engagement estimation.

Furthermore, the current prototype shows the proposed framework's potential for real-time engagement estimation (Figure 3.11). However, the classic machine learning method is lighter. Therefore, we will further investigate the practicality of the proposed framework using a classic machine-learning method. Furthermore, distance learning characteristics and implementation mechanisms, including the possible ethical impacts, must be considered to fully address the practical knowledge gap in an actual distance learning process, which will be discussed in the next chapter. Therefore, the implication and application of these automatic engagement estimation methods will be addressed in the subsequent section.

Chapter 4

System Design of automatic engagement estimation models in distance learning practice

This chapter is an updated and abridged version of the following publications:

1. S. N. Karimah and S. Hasegawa, "MeetmEE: Engagement Estimation-based Online Meeting Room for Distance Learning," in IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE) 2022, 4-7 December 2022,

2. S. N. Karimah and S. Hasegawa, "A Real-time Engagement Assessment for Learner in Asynchronous Distance Learning," in The 17^{th} International Conference on Knowledge, Information and Creativity Support Systems (KICSS), November 23-25, 2022, doi: 10.52731/liir.v003.064,

3. S. N. Karimah, H. Phan, Miftakhurrokhmat and S. Hasegawa, "Design Principle of an Automatic Engagement Estimation System in a Synchronous Distance Learning Practice," in IEEE Access, vol. 12, pp. 25598-25611, 2024, doi: 10.1109/ACCESS.2024.3366552.

4.1 Chapter Introduction

Most existing automatic engagement estimation modules were built from Python, run on a local computer, or developed for robots [111]. To measure cognitive engagement, engagement estimation can be done in parallel with learning activities using LMS log activities for automatic engagement estimation. However, in most cases, the engagement estimation process was done

separately. The learners' video, audio, or physiological cues during learning were recorded, while the estimation process was conducted separately, especially to define emotional and behavioral engagements. Therefore, a real-time accessible engagement report in the actual distance learning process remains unaddressed from the existing studies.

In the previous chapter, we examined two deep learning models for engagement prediction and proposed a framework for a real-time automatic engagement estimation system to analyse learner engagement. A prototype of an automatic engagement estimation tool was developed, based on the proposed framework, partly also to address the RQ3 for helping educators obtain the emotional engagement level records of their learner(s) during the learning process for further evaluation.

This chapter aims at system designs of automatic engagement estimation-based distance learning tools, namely, RAMALAN (a Real-time engAgeMent Assessment for Learner in Asynchronous distaNce learning) and MeetmEE (pronounced as "meet me"). The prototype of RAMALAN and MeetmEE are developed and evaluated by comparing the available feature with the ideal mechanism.

4.2 Distance Learning Characteristics

Due to ICT development and the impact of the COVID-19 pandemic, a paradigm of the learning process has shifted from a traditional classroom to a distance learning system, e.g., massive open online courses (MOOCs) or other online learning activities. As shown in Figure 4.1, in particular, in 2021, 14 articles (29.79%) on the topic of automatic engagement estimation were published (doubled from the previous year) following the outbreak of the COVID-19 pandemic, which started in 2020.

Everybody with technology access can participate and learn anything in distance learning with time and space flexibility. However, distance learning $\neq$ self-learning $\neq$ general online communication. In this thesis, we define that distance learning should at least have: (1) a learner, (2) an educator, (3) learning materials, and (4) an assessment. We use the terms 'learner' and 'educator' instead of 'student' and 'teacher' to note that we aim for distance learning in general and not particularly in a formal education institution.

Based on the interactivity between learner and educator, we characterized distance learning into synchronous and asynchronous distance learning (Figure 4.2). In synchronous distance learning, the learning is conducted through direct communication between a learner and an educator. Video conference applications, such as Zoom, Webex, and Google Meet, are common

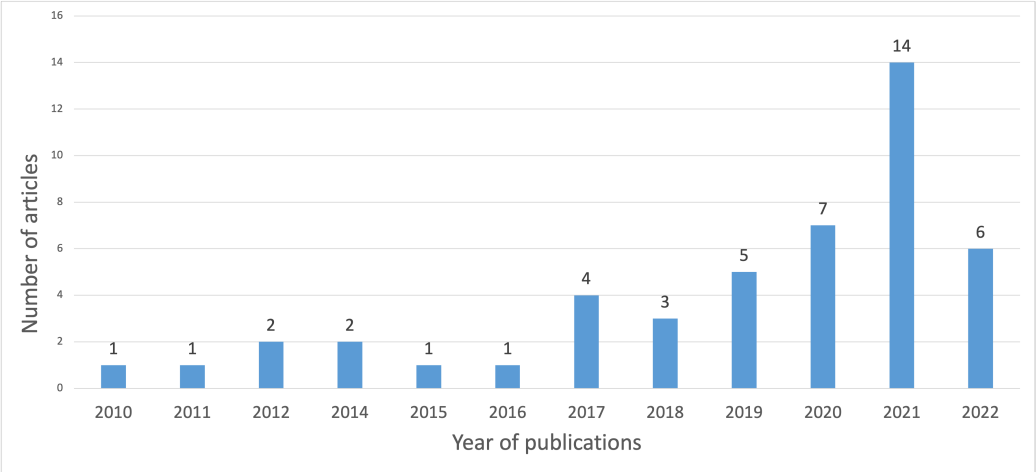


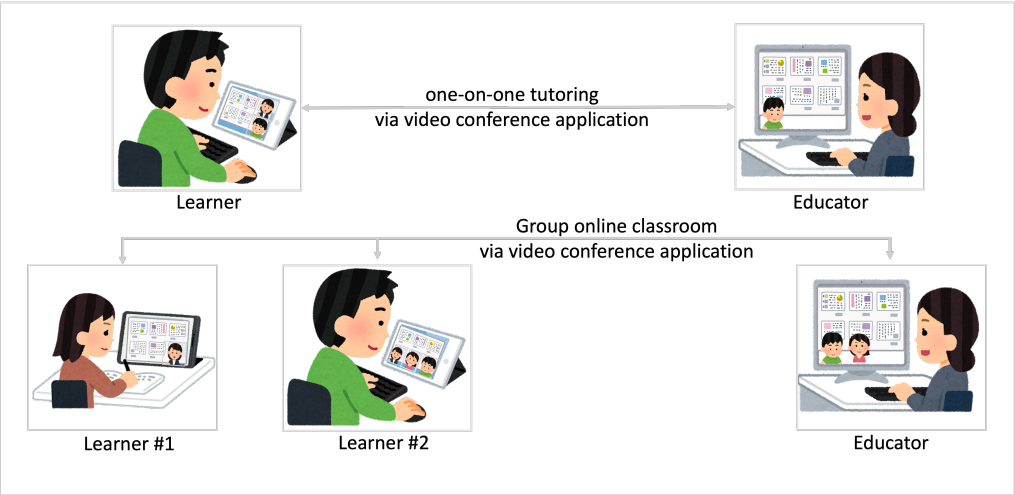
Figure 4.1: Number of reviewed articles per year [109].

applications for one-on-one online meetings in a synchronous distance learning setting. Unlike in synchronous distance learning settings, the learners' visuals in asynchronous distance learning might not be accessible. The learner in asynchronous distance learning is mainly interacting with learning material in a learning portal, massive open online courses (MOOCs), or a learning management system (LMS) (such as Moodle[1]) during the learning. Therefore, educators normally cannot see the learners' faces while interacting with learning material. Thus, measuring their engagement is more complicated than checking their cognitive activity, e.g., by specific tasks, assignments, or exam scores.

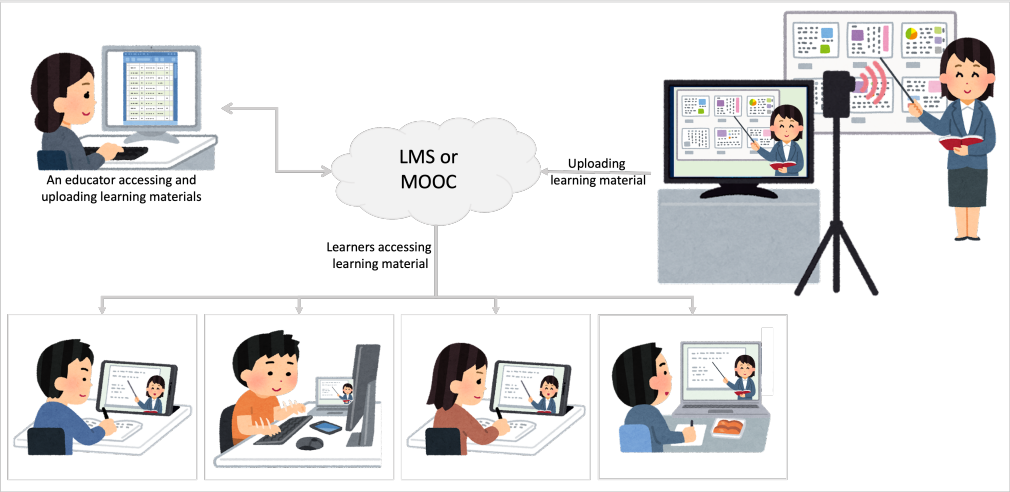
In short, distance learning characteristics can be summarized in 5W1H as follow:

- **What** distance learning should have? a learner, an educator, learning materials, and assessment.
- **Who** can participate in distance learning? Everyone with technology access can be a learner.
- **Why** does a learner participate in distance learning? To learn something or a skill.
- **When** does the learning take place? Anytime/scheduled.
- **Where** does the learning happen? In a place where there is a distance

[1] `https://moodle.org/`



(a) Synchronous distance learning.



(b) Asynchronous distance learning.

Figure 4.2: Synchronous vs asynchronous distance learning.

between the learner and the educator but connected through online communication or learning materials.

- **How** does distance learning conduct? Synchronous or Asynchronous.

4.3 Design System for Asynchronous and Synchronous Distance Learning Tools

With the preliminary knowledge related to distance learning characteristics, we construct systems to integrate the automatic engagement estimation model into distance learning tools. We propose the design of RAMALAN and MeetmEE for asynchronous and synchronous distance learning, respectively.

4.3.1 RAMALAN: a Real-time engAgeMent Assessment for Learner in Asynchronous distaNce learning

A learning management system (LMS), such as Moodle[2], is one practical example of asynchronous distance learning. The LMS has shaped the face of e-learning nowadays since it facilitates many essential educational activities including managing enrollments, creating learning plans, delivering learning content, and grading works in one platform.

In asynchronous LMS, educators normally cannot see the learners' faces while interacting with learning material. Thus, measuring their engagement is difficult other than by checking their cognitive activity, e.g., by certain tasks, assignments, or exam scores. However, the result cannot guarantee that the learners are actually engaged. Besides, due to lack of visibility, the educator cannot check if the learner did the assignment or exam in the LMS by themselves. Therefore, we believe that additional visual information about learners would be beneficial in understanding learners' emotional engagement in asynchronous learning.

Visual analysis through facial recognition is suitable for assessing non-verbal behaviours without interrupting learning. However, unlike in synchronous distance learning settings, where the educator can visually observe learner engagement, a learner is mostly alone in an asynchronous distance learning setting. Therefore, real-time automatic engagement assessment not only benefits educators in adjusting their teaching strategy the way they do in a traditional classroom (e.g., by suggesting some useful reading materials

[2] `https://moodle.org/`

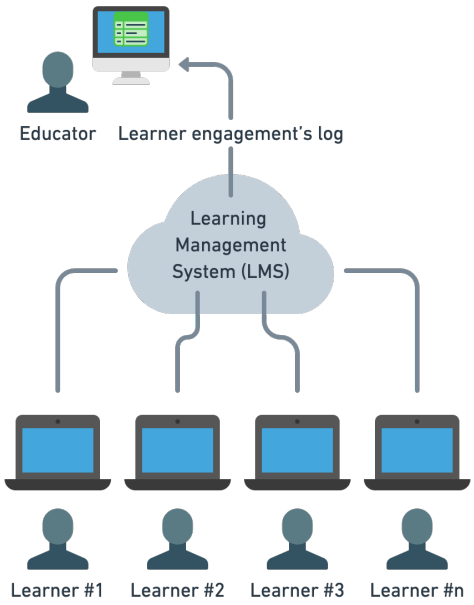


Figure 4.3: Asynchronous learning scenario.

or changing the course contents [229]) but also for self-monitoring by the learner her/himself.

Not to mention the limited bandwidth can be an additional problem for multimedia streaming. Alternatively, a log file that contains log information of learners during their interaction with the learning materials in an LMS can be an option, as shown in Figure 4.3, where the problems should be solved with the following system architecture.

System Architecture

One of the important steps in estimating appearance-based emotional engagement is face detection. Face detection is important to provide the input images and features for the engagement level classification. With the advanced development of artificial intelligence, several real-time facial expression libraries exist, such as OpenCV face rocognition[4], Dlib face recognition[5], and

[4] `https://docs.opencv.org/3.4/da/d60/tutorial_face_main.html`
[5] `http://dlib.net/face_recognition.py.html`

Mediapipe[6]. These libraries help researchers to develop a facial expression-based application. For real-time/online implementation, the face detection application can be deployed in a browser.

In the previous chapter, this dissertation discussed ways to build automatic engagement estimation models using deep learning models, i.e., LSTM and CNN, to predict the emotional engagement of learners in distance learning settings. A real-time automatic engagement estimation system (Figure 4.4) was then proposed using the CNN model, where our LSTM model is not feasible for real-time scenarios due to the feature extraction method.

The initial prototype works well using CNN run in Python Flask. However, a model in JavaScript might be more desirable for web application deployment flexibility in the integration process. Currently, TensorFlow has provided TensorFlow.js, a library for machine learning in JavaScript. It allows machine learning in the browser on the client side, which gives the benefit of higher privacy and lower serving costs. The available pre-made models[3] is easy to use as JavaScript classes. Some pre-made models, such as Pose detection and face landmark detection, can be utilised together to build an engagement estimation model. Alternatively, the existing models can be run by pre-packed them to JavaScript or converted from Python.

While TensorFlow.js can be one potential alternative to develop a distance learning application with real-time engagement estimation, we do not use TensorFlow.js in the early stage of developing the tool; although we might use it for future work. Despite its advantages, we want to further investigate the practicality of our proposed framework in a JavaScript environment web application. Therefore, only the classical machine learning model will be implemented in the proposed framework. Moreover, we decided not to use DAiSEE due to the labeling problem. As shown in Figure 3.2, it is difficult to distinguish between the images with different class labels.

Instead, we defined the poses representing the three engagement levels by obviously visible gestures to build a new dataset. The labeling is mainly based on the distance between the learner's face from the monitor and if the learner is facing the monitor. Very engaged means that the learner continuously faces the monitor closely. Similarly, normally engaged also shows the learner's fully attention to the monitor only with more distance than the very engaged. Meanwhile, not engaged is to classify the learners when their faces were away from the monitor. In this work, we did not take into account more situations, such as note-taking, in which the learner is concentrating by taking a note but looking away from the monitor. Even though facial expressions are not

[6] `https://google.github.io/mediapipe/`

[3] `https://www.tensorflow.org/js/models`

used to define the class label, facial and body features are captured and used for training and prediction.

The system is incorporated with three main steps: data collection, model training, and online implementation. The data were collected by using a face detector to extract the landmarks (face, hand, and body) and the class label for supervised learning. It was packed in an engagement dataset, which is called Engagement.csv in this work, to be trained in some classic machine learning models, such as logistic regression (LR), random forest (RF), and gradient boosting (GB). The trained model aimed to classify three engagement levels: very engaged, normally engaged, and not engaged.

The trained model is saved and implemented in a web-based application that learners can access at the same time during their visit to an LMS. Therefore, their engagement was estimated in real time. The engagement states are automatically recorded in a log file when the estimation process starts to run, analysing learners' faces and body features. The log files were accessible so that the educator could understand the engagement of the learners and give further feedback. In the current architecture, the data is stored independently of the LMS or, possibly, learning record stores (LRS) for simplicity in implementation, and it can be reusable for synchronous architecture as well. However, the current architecture can be further modified to store the data directly in the LMS or LRS for more secure storage.

Prototype Development

The system was implemented in a web-based application with Python run on Flask. We used Mediapipe[1] for the face detector used in both data collection as well as the online running. Before building the web application for real-time estimation, we also first built an application for data collection. We used Mediapipe because it is a community-based open-source work that offers several machine learning solutions, including face detection, face mesh, iris, hands, pose, and holistic. Most importantly, it offers cross-platform, and customisable for live and streaming media, which is suitable for our current work.

Figure 4.5 shows the running application in four states, i.e., the three engagement states plus the idle state, in which the estimation is not running. For ethical reasons, the estimation was not running immediately when the application first ran. Instead, we provided three buttons: start, stop, and capture, which gives the student free will to activate (start button) and deactivate (stop button) the engagement prediction. The capture button

[1] `https://google.github.io/mediapipe/`

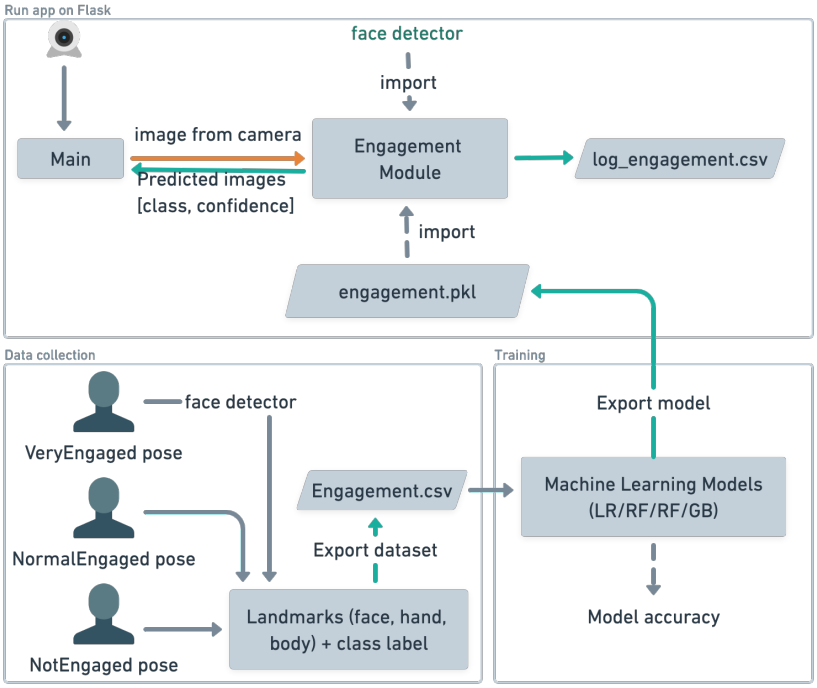
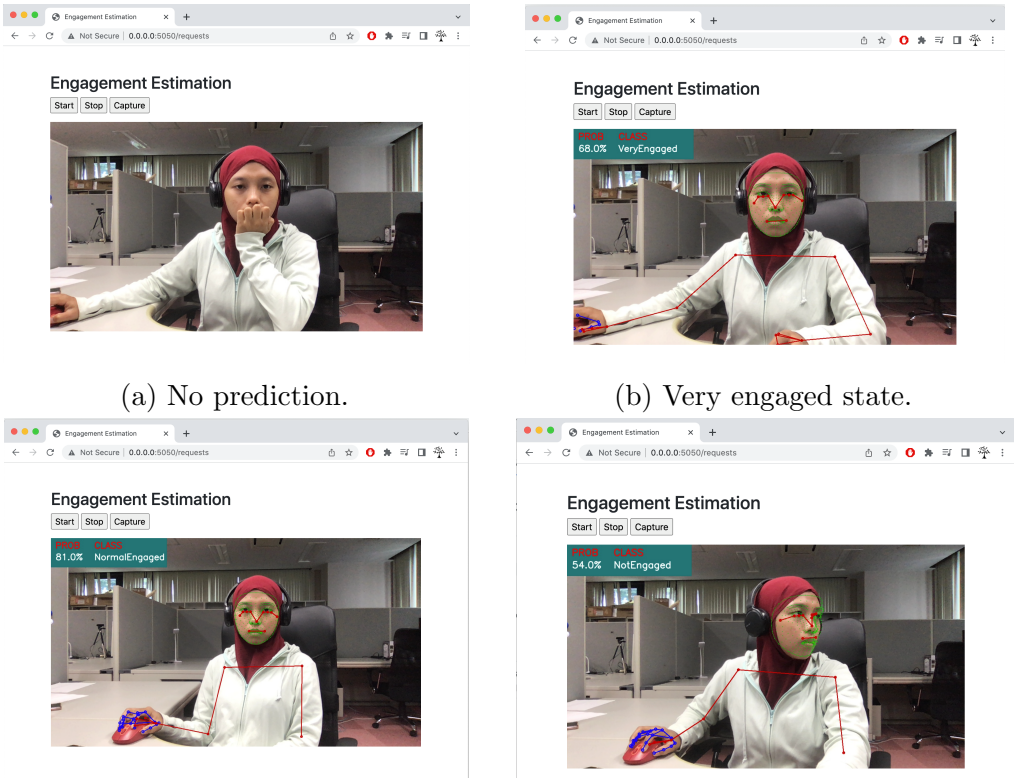


Figure 4.4: The proposed system architecture for asynchronous distance learning.

is additional in case images are needed for further analysis. As a reference, Figure 4.6 shows the screenshot of the engagement log file.

4.3.2 MeetmEE for Synchronous Distance Learning

In a synchronous distance learning setting, the educator can see the learner's face in a traditional classroom. In this learning setting, video conference applications, such as Zoom, Webex, and Google Meet, are common applications used for one-on-one online meetings. In general, estimating a learner's engagement in a synchronous learning setting is not difficult because the learning is mainly conversational, especially, when no learning materials are presented. However, the educator faces challenges tracking the learner's engagement during the meeting due to the difficulty of paying attention to the learner's facial expression [19]. For example, when an educator focuses on giving a lecture through his/her teaching materials, it is difficult to pay attention to the learner's face to assess their engagement simultaneously. Without a real-time engagement estimation system, the subtle changes in engagement



(a) No prediction.

(b) Very engaged state.

(c) Normally engaged state.

(d) Not engaged state.

Figure 4.5: Web-based application of engagement estimation. (4.5a) is the screen in the first load or when the stop button was pressed, whereas (4.5b-4.5d) are the screen views showing the three engagement stages when the start button was pressed.

state in time series are not recorded; therefore, the early states are forgotten. The engagement state log is important as it helps educators monitor and record their learner's engagement for further analysis and to improve the distance learning process.

Current video conference applications do not include automatic engagement estimation. Although many automatic engagement estimation modules have been proposed, most of them were standalone systems that were not integrated with synchronous distance learning practices such as video conference applications. Therefore, we propose the idea of MeetmEE, an engagement estimation-based meeting room for a one-on-one online meeting, e.g., for private tutoring and educator-learner communication. As the initial design of MeetmEE, we focus on one-on-one online meetings in this study.

MeetmEE is designed to enable the educator and the learner to assess

log_engagement-20220825_10

10:51:45	VeryEngaged	70.0%
10:51:46	VeryEngaged	68.0%
10:51:46	VeryEngaged	70.0%
10:51:46	VeryEngaged	70.0%
10:51:46	VeryEngaged	64.0%
10:51:47	VeryEngaged	70.0%
10:51:47	VeryEngaged	66.0%
10:51:47	VeryEngaged	68.0%
10:51:48	VeryEngaged	68.0%
10:51:48	VeryEngaged	75.0%
10:51:48	VeryEngaged	76.0%
10:51:49	VeryEngaged	75.0%
10:51:49	VeryEngaged	77.0%
10:51:49	VeryEngaged	67.0%
10:51:49	VeryEngaged	66.0%
10:51:50	NormalEngaged	82.0%
10:51:50	NormalEngaged	82.0%
10:51:50	NormalEngaged	87.0%
10:51:51	NormalEngaged	90.0%
10:51:51	NormalEngaged	84.0%
10:51:51	NormalEngaged	89.0%
10:51:52	NormalEngaged	86.0%
10:51:52	NormalEngaged	88.0%
10:51:52	NotEngaged	76.0%
10:51:52	NotEngaged	91.0%

Figure 4.6: A snapshot of the automatic engagement log in a CSV file.

their engagement during the meeting and retrospectively evaluate distance learning. We utilized: (1) Web real-time communication (WebRTC) to construct peer-to-peer audio/video communication, (2) facial recognition and engagement estimation modules, and (3) a live diagram plot to show the real-time engagement state graph. An engagement log can be downloaded for further analysis.

Web Real-Time Communications (WebRTC)

Several video conferencing systems have been described in the literature, including peer-to-peer (P2P) [251, 196, 70] and star topology [222]. WebRTC can be integrated with an existing distance learning system to enhance distance learning implementation [8], such as an additional feature for video conferencing in a learning management system (LMS) [211]. However, the WebRTC in the existing application merely served a single purpose, i.e.,

audio/video/data transmission, instead of providing additional features. By adding face and object detections in WebRTC implementation [163, 105], the system has more functionality, such as enabling a remote image collection to create an initial database for enhancing face recognition [104].

WebRTC is a standard that includes protocols and JavaScript APIs to enable real-time communication among web browsers [102, 132]. WebRTC works by defining an API that allows browsers and scripting languages to interact with media devices (microphones, webcams, and speakers), processing devices (encoders/decoders), and transmission functions [132]. The architecture of WebRTC includes end-user clients, back-end media components (for example, a WebRTC media broadcasting server), and signaling servers [173].

WebRTC utilizes three primary APIs of a browser, i.e., MediaStream for acquiring audio and video streams, RTCPeerConnection for communication of audio and video data, and RTCDataChannel for communication of arbitrary application data [102]. Besides APIs, signaling plays a significant role in WebRTC implementation. A signaling protocol employs Interactivity Connection Establishment (ICE), Session Traversal Utilities for NAT[1] (STUN), and Traversal Using Relays around NAT (TURN), to exchange session descriptions in the form of Session Description Protocol (SDP) [132, 102].

Open-access signaling server libraries and SDK, such as socket.IO[2] and SkyWay[3], are available to develop the communication between a client and a server in WebRTC. Socket.IO is a library built on top of the WebSocket protocol. Its server implementations are available for Javascript, Java, Python, and Golang, whereas the client implementations are available in most major languages, such as Javascript, Java, C++, and Python [**?**]. SkyWay (ECLWebRTC) provides SDK and API for easy implementation of WebRTC. It also provides all the required servers for WebRTC, including signaling servers, STUN, TURN, and SFU servers. The SDKs are available in multi-platform such as Javascript, iOS, and Android. In addition, peer authentication is available to prevent billing problems caused by the unauthorized use of API keys. Moreover, SkyWay also provides a WebRTC gateway for more implementation on IoT devices (such as Raspberry Pi) and game engines (such as unity)[**?**].

Skyway is a quick and easy solution for developing a WebRTC. However, registration is required to use the service. Therefore, we used socket.IO in this work because it enables bidirectional communications without registration and provides more room for experimentation and modification.

[1]Network Address Translation

[2]`https://socket.io/`

[3]`https://webrtc.ecl.ntt.com/en/skyway/overview.html`

System Architecture and Protoype development

To develop MeetmEE, we integrate web real-time communication (WebRTC) and engagement estimation modules as shown in Figure 4.7, where the data collection and engagement training model uses the model as the one used in RAMALAN.

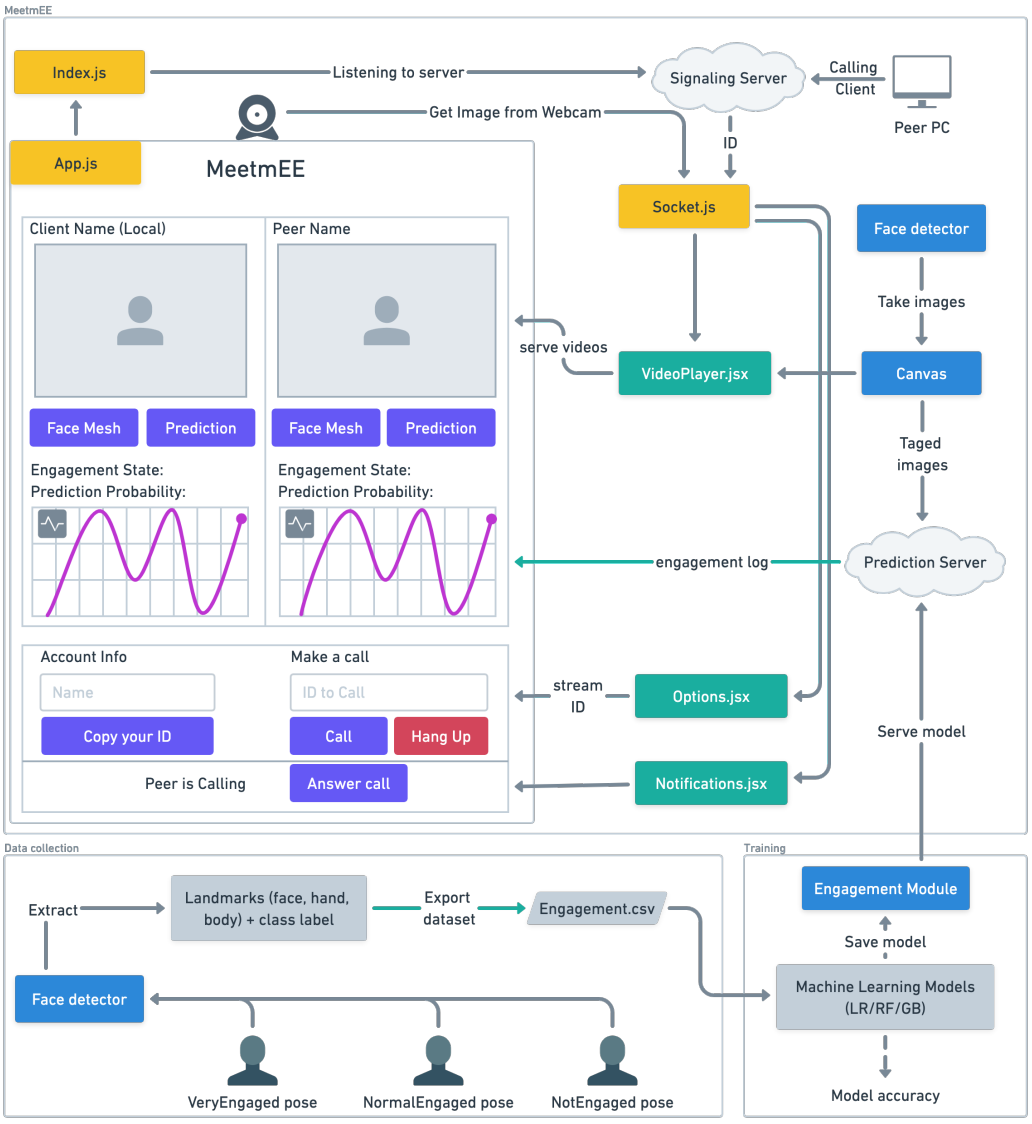


Figure 4.7: MeetmEE Architecture for synchronous distance learning.

The proposed system design comprises three major components, i.e., output, communication, and engagement modules, represented in green, yellow, and blue color boxes in Figure 4.7. The output component consists of

modules to serve the video streams (from both peers), the facial landmark drawings, the engagement predictions and graphs, and other menus shown on the front end. The communication component plays a key role in WebRTC communications, including giving ID for both peers, streaming the audio/video data, and call/answer/hang up functions. Likewise, the engagement modules consist of facial recognition and engagement modules; the output component will serve as the result on the page. The engagement estimation log can also be downloaded from the main page for further analysis. Figure 4.8 shows the current development of the MeetmEE prototype.

One of the constraints of the proposed system design is that it currently works only for one-on-one meetings. Improving the scalability of the system design to enable multiple participants (e.g., an online classroom) would improve its practicality in synchronous distance learning practices. Moreover, the current work focuses on defining the engagement level by frontal images and the distance of the participant from the monitor. Therefore, some behaviours during the meeting, where the face is directed away from the monitor, such as writing or doing other related assignments, would be estimated as low engagement. Including more engagement cues, such as log files and physiological cues, in addition to frontal-images analysis will improve the estimation accuracy.

Furthermore, there is a possibility that this technology could easily be abused from an ethical impact perspective. Therefore, developing the proposed design system should include message encryption, authentication and access control, and automated data expiry rules to mitigate the potential abuse of the technology used. In addition, the implementation should also follow the mechanism in Figure 5.1 that addresses technical and operational measures.

4.4 User experience evaluation

The evaluation phase aims to collect feedback regarding the user perspective of automatic engagement estimation implementation. In this work, MeetmEE is deployed as a representative implementation to understand the user perspective in actual distance learning practise.

The current prototypes show how the emotional engagement of both the learner and educator is recorded during one-on-one synchronous learning, where the engagement module can be added to the present common video conference application (such as Zoom, Webex, Google Meet, etc.) for full implementation.

In this pilot experiment, we evaluate the usability of the MeetmEE system and the uncertainty that users face in the automatic engagement estimation

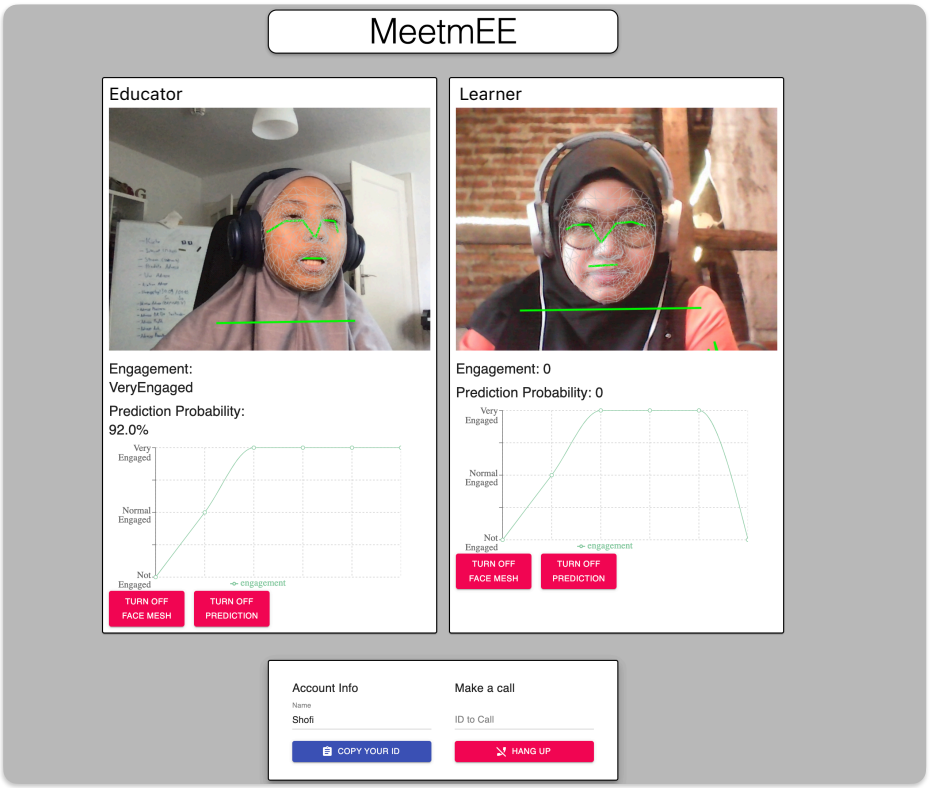
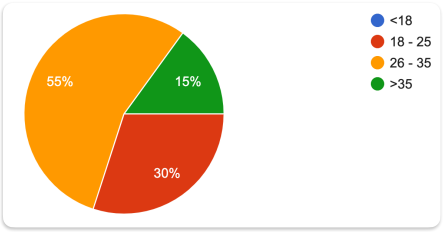


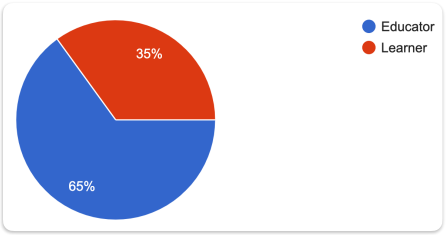
Figure 4.8: Screenshot of the MeetmEE prototype from educator's side, where both face mesh and prediction buttons are on

implementation concerning engagement estimation results, data security, or privacy.

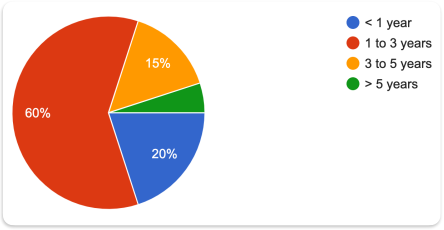
A total of 20 participants joined the experiment either as educators ($n = 13; 65\%$) or learners ($n = 7; 35\%$), with 60% of them participating in distance learning for one to three years. The participants completed the survey based on their roles in their affiliations in Indonesia, Japan, or Taiwan. Note that the evaluation is conducted fully online to reach the participants in Indonesia and Japan. Figure 4.9 shows the details of the participation profiles.



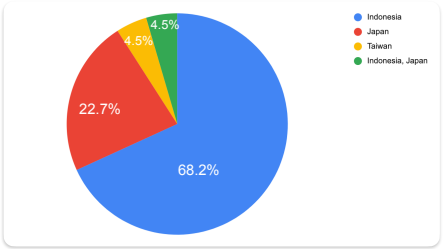
(a) Age.



(b) Role in distance learning.



(c) Distance Learning Experience.



(d) Country of affiliation[a].

[a]One participant filled 2 affiliations.

Figure 4.9: Profiles of the participants.

4.4.1 Experiment settings

The term experiment refers to a pilot user experience experiment conducted as a one-on-one online meeting in MeetmEE. Depending on the role, either as an educator or a learner, each participant joined a one-hour meeting session with the author via MeetmEE (Figure 4.10). During the session, the participants discussed their experience in distance learning with the author while filling out two survey forms, Form A and Form B. Form A focuses on evaluating the idea and system design, whereas Form B focuses on the technological evaluation through the user experience. Each participant provided informed consent before the experiment.

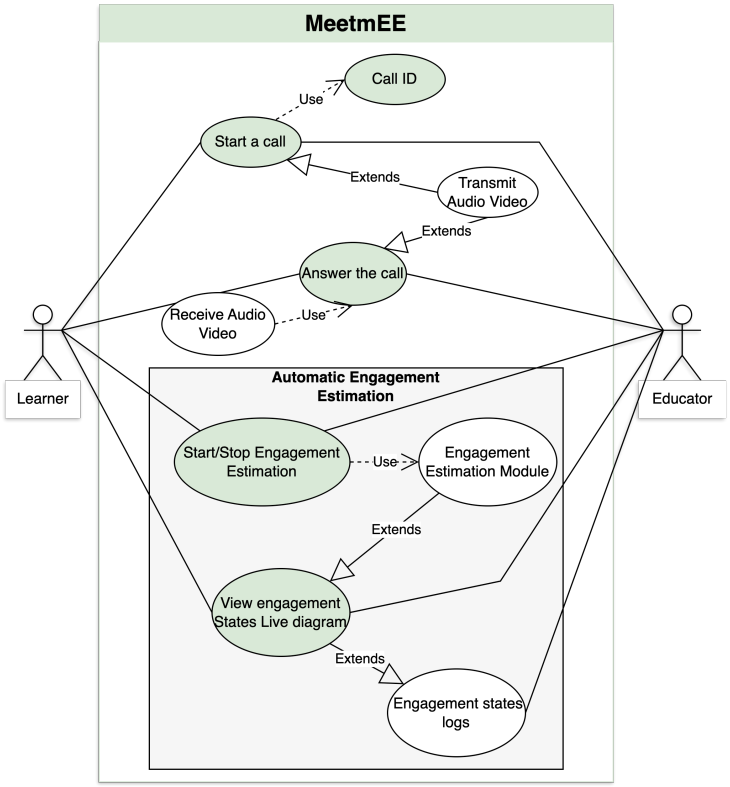


Figure 4.10: MeetmEE use case on the experiment.

Form A

Form A is intended to evaluate the idea and users' perspective on the implementation of automatic engagement estimation in distance learning practice. The questions on Form A were formed to understand how engagement estimation in synchronous distance learning would help educators improve their teaching methods and learners become more engaged in learning. Extending the survey on [120], we present the most remarkable questions as shown in Table 4.1.

Except for Q1, the most remarkable questions of Form A are answered with a five-point Likert scale: strongly disagree (1), disagree (2), neutral (neither agree nor disagree) (3), agree (4), and strongly agree (5).

From B

To evaluate the user experience on the MeetmEE interface, we utilised the User Experience Questionnaire (UEQ) tool[187], which interprets the items in the questionnaire into six scales:

1. Attractiveness: Overall impression of MeetmEE, whether users like or dislike.

2. Perspicuity: Whether it is easy to get familiar with and learn to use MeetmEE.

3. Efficiency: Whether the users can solve their tasks without unnecessary effort.

4. Dependability: Whether the user feels in control of the interaction.

5. Stimulation: Whether it is exciting and motivating to use MeetmEE.

6. Novelty: Whether the product is innovative and creative as well as catches users' interest.

Table 4.2 shows 26 pairs of contrast items to construct UEQ scales. Each pair of contrast items is represented in 1 to 7 values, where 1 is on the extreme left item and 7 is on the extreme right item. Note that the order of the positive and negative items is randomized in the questionnaire. Not that UEQ does not produce an overall score for the user experience, which can be done using the KPI extension.

Table 4.1: The most remarkable questions of Form A.

Q Nr	General questions for educators and learners
1	I know my peers are engaged with the distance learning/meeting from..... (The facial expression and body language / The active participation in the class / The assignments and scores)
2	I notice any changes in my emotions and my peer/student's emotions in any distance learning process.
3	It is appropriate and ethically fair to monitor learners' engagement from their facial features in the distance learning process.
4	To understand learners' emotional engagement, automatic engagement estimation should be implemented in distance learning tools such as learning management systems (LMS), and online classrooms through video conference apps (e.g., Zoom, Google Meet, Webex).
5	I feel that automatically estimating emotional engagement is a reasonable and appropriate feature in distance learning practice.
6	MeetmEE is easy to use.
7	MeetmEE will make the lecture more interesting.
8	MeetmEE is practical.
9	I think that MeetmEE is a reasonable and appropriate feature for automatically estimating emotional engagement in distance learning practice.
10	MeetmEE would be a welcome addition to a lecture.
11	I notice any changes in my emotions and my peer's emotions as a result of MeetmEE's emotion estimation.
12	MeetmEE is a potential feature to enhance distance learning.
Q Nr	Educator questions
1	I was aware and comfortable with automatic engagement estimation recording my engagement state from my face and body looks while doing the learning activities.
2	Having learner engagement records from MeetmEE gives me insights to understand learner engagement in the entire learning session.
3	I think monitoring learners' engagement will disturb learners' activities in distance learning.
4	The use of this technology will motivate me to improve my teaching strategies and give personal support to my students.
5	Therefore, I would like to use MeetmEE to enhance my teaching.
Q Nr	Learner questions
1	I was aware and comfortable with automatic engagement estimation recording my engagement state from my face and body looks while doing the learning activities.
2	If I could choose, I would let MeetmEE capture my emotional engagement states during learning activities so my teacher would know when I show a disengagement sign and give me personal support or change his/her teaching strategy.
3	MeetmEE enabled me to measure my own engagement.
4	MeetmEE stimulated my motivation to keep engaged during the learning session.
5	MeetmEE made the learning interesting to me.
6	MeetmEE distracted me from learning.

Table 4.2: UEQ Scale[187]

Nr	Item	Scale
1	annoying/enjoyable	Attractiveness
2	bad/good	
3	unlikable/pleasing	
4	unpleasant/pleasant	
5	unattractive/attractive	
6	unfriendly/friendly	
7	(not) understandable	Perspicuity
8	difficult to learn/easy to learn	
9	complicated/easy	
10	confusing/clear	
11	inferior/valuable	Stimulation
12	boring/exciting	
13	not interesting/interesting	
14	demotivating/motivating	
15	unpredictable/predictable	Dependability
16	obstructive/supportive	
17	not secure/secure	
18	(does not) meet expectations	
19	slow/fast	Efficiency
20	inefficient/efficient	
21	impractical/practical	
22	cluttered/organized	
23	dull/creative	Novelty
24	conventional/inventive	
25	usual/leading edge	
26	conservative/innovative	

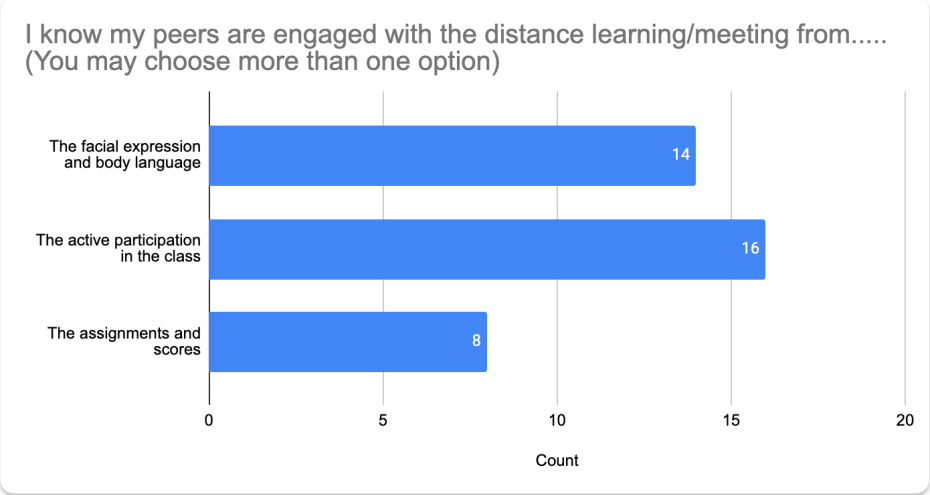


Figure 4.11: Survey results of Q1 general questions

4.4.2 Experiment Results

The results of Q1 (Figure 4.11) demonstrate that most of the participants knew the engagement of the other participants in an online meeting from active participation in the class (behavioral engagement), facial expressions and body language (emotional engagement), and assignments and scores (cognitive engagement). Furthermore, Table 4.3 analyses the results of Form A individually, whereas Figure 4.13 and 4.12 show the results of Form B after value transformation.

Form A results

Based on the results in Table 4.3 and Figure 4.13, most of the responses were very positive to the concept of automatic engagement estimation, which is represented in MeetmEE. In the general questions, although 60% of the participants claimed that they could notice any changes in their emotions and their peers, 30% claimed otherwise (Q2). However, most participants agreed, with an average of 4.05 points on a five-point Likert scale, that automatic engagement estimation technology should be implemented in distance learning practices to understand learners' emotional engagement (Q4). In particular, the participants noticed any changes in their emotions and their peers' emotions as a result of MeetmEE's estimation, in which participants agreed with an average of 3.9 points on a five-point Likert scale on Q11.

Table 4.3: Form A results

Q Nr	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Average points
	General questions results (%, $n = 20$)					
2	20.00	10.00	10.00	45.00	15.00	3.25
3	0.00	0.00	20.00	55.00	25.00	4.05
4	0.00	10.00	20.00	25.00	45.00	4.05
5	0.00	5.00	10.00	30.00	55.00	4.35
6	5.00	15.00	25.00	15.00	40.00	3.70
7	0.00	15.00	20.00	40.00	25.00	3.75
8	0.00	10.00	0.00	55.00	35.00	4.15
9	0.00	20.00	0.00	45.00	35.00	3.95
10	0.00	15.00	15.00	30.00	40.00	3.95
11	0.00	15.00	5.00	55.00	25.00	3.90
12	0.00	5.00	0.00	50.00	45.00	4.35
	Educators questions results (%, $n = 13$)					
1	0.00	30.76	0.00	38.46	30.76	3.69
2	0.00	7.69	7.69	61.54	23.08	4.00
3	15.38	30.76	23.08	23.08	7.69	2.77
4	0.00	0.00	7.69	30.76	61.54	4.54
5	0.00	15.38	15.38	23.08	46.15	4.00
	Learners questions results (%, $n = 7$)					
1	0.00	14.28	0	28.57	57.14	4.28
2	0.00	0.00	14.28	28.57	57.14	4.28
3	0.00	0.00	28.57	42.86	28.57	4.00
4	14.28	0.00	28.57	14.28	42.86	3.71
5	0.00	28.57	14.28	28.57	28.57	3.57
6	14.28	28.57	14.28	14.28	28.57	3.14

Moreover, MeetmEE is a potential feature for enhancing distance learning (Q12), in which the participants agree on an average of 4.35 points on a five-point Likert scale.

In general questions Q3 and Q5, the participants were asked whether an automatic engagement estimation system is appropriate and ethically fair to monitor learners' engagement from their facial features; thus, it is a reasonable feature in the distance learning process. The participants gave an average of 4.05 and 4.35 points on a five-point Likert scale, respectively.

The reason why MeetmEE as an implementation of the automatic engagement estimation feature in distance learning is so favourable for the 70% of the participants is that, for educators, this technology will motivate them to improve their teaching strategies and give support to their students, while students can measure their own engagement as well. Most importantly, the participants, either educators or learners, were aware of and comfortable with automatic engagement estimation recording their engagement state from their face and body looks while performing the learning activities. Note that in MeetmEE, participants can see each other's engagement state, enabling the learner to see the educator's engagement state. An educator participant responded positively to this because she did not feel ignored when her students tracked her engagement during her teaching.

However, despite MeetmEE's capability, the participants responded variably regarding whether implementing this technology disturbs learning activities in distance learning settings. The results Q3 for educators and Q6 for learners suggest that potential disturbance caused by meetmEE showed an almost equal agreement distribution.

Form B results

The overall result of the questionnaire in Form B is depicted in Figure 4.12, in which the scale was transformed into a -2 to +2 scale for better readability interpretation [187]. The results showed a positive evaluation, demonstrating that MeetmEE is sufficient, particularly in scales of stimulation, attractiveness, perspicuity, and novelty.

However, MeetmEE is perceived as relatively low in terms of dependability and efficiency, the details of which are shown in Figure 4.13. The figure shows that most participants experienced MeetmEE as slow and relatively insecure, leading to low efficiency and dependability.

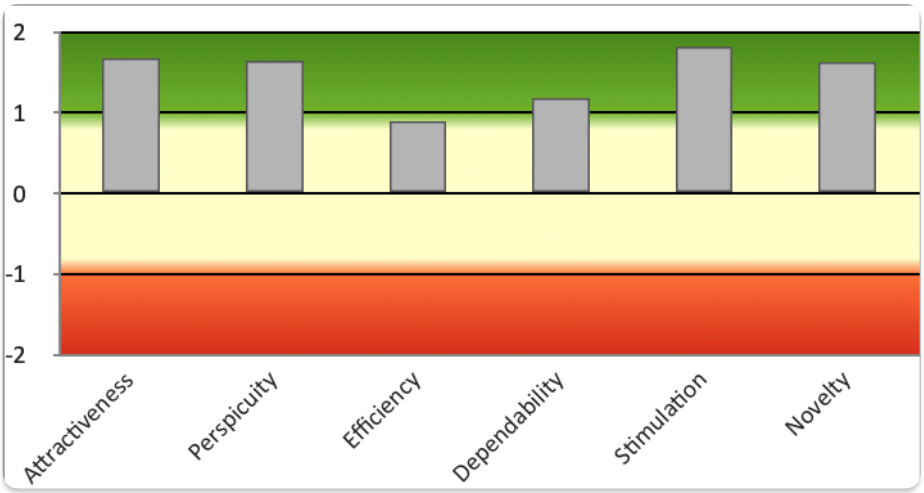


Figure 4.12: UEQ Scale results (mean and variance).

Figure 4.13: UEQ mean value per item after transformation

4.5 Chapter conclusion

In this chapter, distance learning characteristics were discussed to clarify the learning setting aimed, i.e., synchronous and asynchronous distance learning.

We propose the design of RAMALAN (Figure 4.4) for a distance learning tool with integrated engagement estimation technology while considering the ethical impact. We integrate the automatic engagement estimation modules in LMS for real-time engagement assessment in an asynchronous distance learning setting and in WebRTC as a one-on-one online meeting platform for a synchronous distance learning setting.

We presented the problem definition (Figure 4.3 as a simplification image of Figure 4.2b) to give an overview of the scenario where learners, which mainly interact with the learning materials in LMS, had no direct interaction with the educator. Our proposed solution (Figure 4.4) enabled the educator to obtain engagement state logs of their student in less-bandwidth-demand form.

Although the images and videos of learners will not be recorded, our proposed system analysed emotional engagement, where visual information of face and body were extracted. Therefore, we provided the application with start/stop buttons so they are aware of when their faces will be analysed (Figure 4.5).

We also propose MeetmEE, a potential means of enhancing synchronous distance learning practices through a one-on-one online meeting with engagement estimation. The automatic engagement estimation modules in the system consisted of a face detection module and an engagement recognition module to estimate the engagement level during the meeting and record it as a downloadable log file.

A pilot experiment was conducted as a user experience survey to evaluate the MeetmEE system design and construct the ethical implementation design principle. A total of 20 participants joined the experiment in a one-hour meeting session with the author via MeetmEE online either as educators ($n = 13; 65\%$) or learners ($n = 7; 35\%$) with 60%. The participants completed two survey forms (Forms A and B) based on their roles in their affiliations. The experiment results of Form A demonstrate that most of the responses were very positive to the automatic engagement estimation concept, represented in MeetmEE.

MeetmEE is favourable for 70% of the participants, where, for educators, this technology will motivate them to improve their teaching strategies and give support to their students, while students can measure their own engagement as well. Furthermore, the results of Form B showed a positive evaluation, demonstrating that MeetmEE is sufficient, particularly in scales

of stimulation, attractiveness, perspicuity, and novelty. However, most participants experienced MeetmEE as slow and relatively insecure, leading to low efficiency and dependability, and therefore, leading to potential abuse of an ethical perspective.

Despite the merits of the proposed designs, the RAMALAN has not been tested in actual LMS for real implementation in an educational learning process. Meanwhile, the MeetmEE implementation is lacking in scalability, stability, and estimation reliability. Furthermore, although the ethical issues have been taken into consideration in the current system designs, i.e., adding control buttons for prediction, more measures are required to ensure the ethical implementation, which will be discussed in the subsequent chapter. Meanwhile, the ethical remain unaddressed in the existing works (Chapter 2).

Chapter 5

Design Principle of an Automatic Engagement Estimation System in a Synchronous Distance Learning Practice

This chapter is an updated and abridged version of the following publications:

1. S. N. Karimah, H. Phan, Miftakhurrokhmat and S. Hasegawa, "Design Principle of an Automatic Engagement Estimation System in a Synchronous Distance Learning Practice," in IEEE Access, vol. 12, pp. 25598-25611, 2024, doi: 10.1109/ACCESS.2024.3366552.

5.1 Chapter Introduction

Although the participants in the user experience experiments agreed that an automatic engagement estimation system is appropriate, implementing this technology in distance learning settings could easily be abused from an ethical impact perspective. Not to mention whether the application is not secure. Therefore, ethical procedures should be considered when deploying the proposed design system to mitigate potential abuse of technology. This chapter discusses the ethical risks pertinent to privacy protection. Furthermore, the design principle of the automatic engagement estimation implementation is proposed.

Furthermore, the design principle of the automatic engagement estimation implementation is proposed.

5.2 Ethical Risks Pertinent to Privacy Protection

The four ethical values and issues principles are autonomy, non-maleficence, beneficence, and justice [207]. Autonomy, which is perceived as the ability to make own decisions, and beneficence, which concerns the positive impact of an act, have implications regarding privacy rights [103]. Ethical violations are present, such as privacy infringement, disclosure of identifiable individuals to the public, or data misuse [103].

Ethically, the information owners must provide informed consent before any party can legally use their data. They have the right to access that information, correct it, and request no further data to be collected [114]. Informed consent is given with the knowledge of all the facts needed to make a rational decision [114]. However, if a person is not aware of the basic concept of information security, s/he is more prone to information security threats than others [118]. In addition, many services and ICT applications collect users' consent in absolute terms of use, which leaves no option for people to use the service or application without agreeing to all terms [183, 250, 103]. Such consent is ethically dubious and inadequate for protecting privacy.

In distance learning terms, users as information owners, especially learners, are the aggrieved party if their affective information is leaked. Therefore, in addition to information security awareness, the implementation of automatic engagement estimation in the distance learning process should provide a built-in information security mechanism to ensure that information is not accessed, used, disclosed, recorded, or modified by unauthorized entities [114, 75]. Before proposing a safe implementation mechanism, discussing possible ethical issues resulting from inadequate privacy protection is important to understand ethical standardization.

Failing to develop ethical standards and procedures in a distance learning environment minimises the effectiveness of automatic engagement technology and therefore decreases the value perception of the system [4]. The following subsections outline three ethical issues that will be conflicted in a haphazard implementation of automatic engagement estimation in distance learning practice: data misuse, undermining trust and learning mood, and reluctance to join the learning.

5.2.1 Data misuse

A primary ethical concern for any system that processes personal data is that it may be misused in such a way as to cause harm to data owners or

their property [103]. Educational technology systems could misuse this by educational institutions (e.g., school principals), educators, or others with access to data, such as operators.

Appropriately, the principals use learners' data mostly to evaluate the institution, make improvements, and model best practices of data use, while educators use the data to improve instruction and outcomes for the learners [142]. However, some actions are ethically in the grey area. For example, educators use learners' data to communicate with colleagues outside the content area or to initiate conversations and support collaboration among educators [142].

Should a data specialist, a technology assistant, or an external operator outside the educational institution and the responsible educators gain access to the data, they could engage in various data misuses. For example, a data specialist or any other external party could use learners' affective data for another research purpose outside the agreed terms with the educational institution or even learners. Affective data can also be used for deep-face databases that are not related to distance learning purposes in this context. Therefore, technical and operational security measures should be implemented to reduce the risk of data misuse by restricting unauthorized parties from accessing system data [103].

5.2.2 Undermining trust

Trust is a major concern in distance learning [4]. For learners, the main concern is trust in the learning system they use as well as the protection of their sensitive personal information [4, 12]. Implementing an automatic engagement estimation without a privacy plan or consideration leads to learners' perception of being unethically monitored and observed. Therefore, in addition to data misuse, there is a risk of undermining learners' trust in using distance learning tools integrated with automatic engagement estimation.

5.2.3 Reluctance to participate in distance learning

One ethical aspect of e-learning, including distance learning, is the motivation to use ICT tools in education [190]. Many learners and educators are not motivated to effectively use new technology in learning and teaching because of factors such as satisfaction with the tool, interest level, and joy when using the tool [190, 191]. Moreover, failing to establish trust between the users involved in the system will minimize the value perception of distance learning systems [4]. Furthermore, flexibility is one benefit that a distance learning setting offers learners participation in a learning process, even in private

time and place. However, participating in learning in private time and places causes inconvenience for learners when turning on their cameras, especially in synchronous distance learning. In addition, some learners merely do not like to show their faces either because of shyness or privacy concerns, which hinders the assessment of learner engagement using visual cues. Therefore, the forced implementation of automatic engagement estimation leads to learning demotivation and reluctance to participate, which is contra-productive for assessing engagement in distance learning, such as reducing dropout.

5.3 Design principle of ethical engagement estimation technology implementation in distance learning process

To implement an enhanced distance learning tool while considering the ethical impact, we introduced a design principle for the ethical implementation of automatic engagement estimation in distance learning practice (Figure 5.1). Protecting privacy in an online network requires cooperation from both the control authority and individual users [248]. Therefore, the design principle involves technical and operational measures that include the participation of both authorities (e.g., education institutions) and individual users (e.g., learners and educators) while attempting to mitigate the aforementioned risks.

5.3.1 Technical Measures

Technical measures to mitigate the risks include trustworthy architecture, estimation models, and data security.

Trustworthy architecture

A trustworthy architecture for automatic engagement estimation leads to a good user experience, especially in terms of attractiveness, perspicuity, and efficiency. However, a non-trustworthy architecture potentially causes poor development and performance, which leads to reluctance to use the feature in the distance learning process.

For instance, in the pilot experiment, MeetmEE obtained a good user perception in terms of attractiveness and perspicuity, yet low efficiency owing to its slow performance. As shown in the MeetmEE architecture (Figure 4.7), the face detector takes and tags the images from the streamed video and then

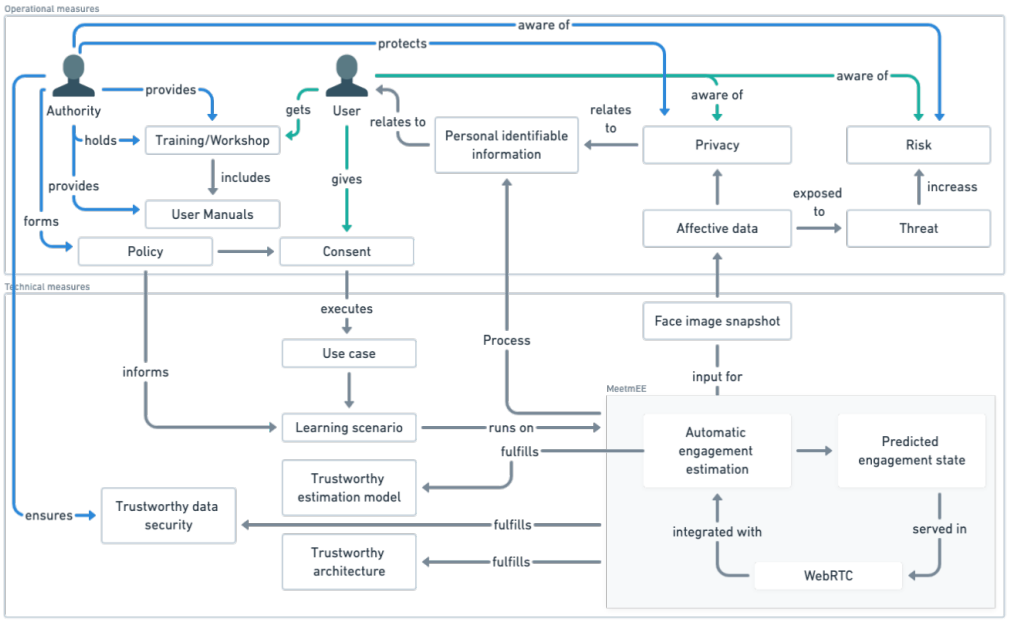


Figure 5.1: Design principle of automatic engagement estimation implementation in distance learning settings. The user in the figure can be the educator or learner.

sends the tagged images to the prediction server. In this architecture, the latency from capturing the first snapshot to be predicted and serving the result back over the online network was high, even though only plain images were processed. This process causes the application to operate slowly when the prediction feature is on.

Therefore, the MeetmEE system design provides two control buttons for the face mesh and prediction to give the user awareness of the prediction feature and technical stability. The future implementation of automatic engagement estimation should consider a system design architecture with alternative input for the prediction server while maintaining good performance.

Trustworthy estimation model

Once an automatic engagement estimation model is prepared and deployed as a feature in a synchronous distance learning setting, the model outcomes (i.e., the predicted engagement states) must be regularly monitored and measured to gain insight into the learning-teaching experience. The trustworthy estimation model addresses key issues regarding model quality metrics, model fairness, drift detection, and explainability [69].

Model quality metrics Model quality metrics such as accuracy, precision, and areas under the ROC, etc. measure the model quality during the entire lifecycle and allow corrective actions to be taken in model development [111, 69].

Model fairness Model fairness refers to bias related to preferences for certain values of the chosen features such as gender, age, ethnicity, nationality, etc. [69]. The automatic engagement estimation model in the MeetmEE prototype was built based on the Asian facial type and tested on specific user profiles (Figure 4.9). Databases from different cultural and ethnic backgrounds may require more adjustments for model development [110, 111].

Drift detection Drift detection measures accuracy and prediction consistency. We observed that the predicted engagement level decreased and was inconsistent when the MeetmEE ran online. The drop in accuracy is conceivably due to the high latency in the client-server processing, which refers to the front-end application to the prediction server. Meanwhile, the inconsistency is owing to false class encoding on the front-end side.

Explainability Explainability provides insight and transparency of the engagement model outcomes to educators with no data science skills. The live graph under the video stream on MeetmEE shows the real-time engagement state as an outcome of the engagement estimation model.

Trustworthy data security

In a synchronous learning session, the audio-video stream data are the main inputs for automatic engagement estimation. In MeetmEE, the face detector takes and tags the images from the streamed video and sends the tagged images to the prediction server. One prediction cycle in MeetmEE incorporates plain images without encryption or security measures.

Ideally, the data sent between the educator and learner in the run system should be encrypted to avoid the risk of third-party interception if the messages contain potentially identifiable data. An encoder-decoder mechanism is desirable in the application of automatic engagement estimation. For example, when the system runs, only the extracted features or compressed data at a certain period are shared under appropriate encryption. A conceivable system should be designed with a decoder to handle encrypted data. Therefore, without the correct decoder, the data would be meaningless. Encrypted

data should include messages as well as any interactions, such as the estimated engagement levels. However, applying secure end-to-end integration of automatic engagement-estimation applications remains challenging.

Furthermore, a secure database implementation is suggested to defend against hacking or any other unauthorised access [103]. Additional software measures include message encryption, authentication and access control, automated data expiry rules, log reports, and data anonymisation [103, 151]. The security mechanism can also include how long the data should be archived (including backups) and the measures to delete the content [4].

5.3.2 Operational Measures

Operational measures are required in the implementation of automatic engagement estimation because mere technological tools are insufficient to protect information security, in which the human factor remains a major vulnerability [103]. An operational efficacy approach that supports effective human decision-making is necessary to mitigate risks and end-user behaviour. The following operational measures are recommended to ensure the ethically safe implementation of automatic engagement estimation in distance learning:

Self awareness Both authorities and users of automatic engagement estimation should protect learners' personal information. Likewise, users should be aware of their personal information and provide their consent. Furthermore, both authorities and users should be aware of the threats and risks of data exposure.

In the MeetmEE interface, the on/off buttons for face mess and prediction enable the user to be aware of the automatic engagement feature that predicts their engagement during the learning session. Users have free will to activate or deactivate the feature so that no ethical issues are violated.

Fair consent Authorities should form the policy and provide fair consent following ethical standards and laws. Fair consent is one approach to list a policy framework that should encourage user awareness to protect the data they would or would not provide. However, a major problem in most privacy policy agreements regarding the use of technology is the presence of NO-OPTION. For example, after the long privacy policy points, there is only one option: to agree. Otherwise, the use of the technology is impossible. Therefore, policy consent should be written fairly, considering the main goal of distance learning. In the case of children, they must be accompanied by their parents or responsible adults to provide consent.

Multiple options should be available in the implemented system so that both the educator and learner benefit from automatic engagement estimation in the distance learning process. A good distance learning system should allow the learner to choose what data are shared (video stream or engagement log), and what data to show. For example, if a learner does not want to show their face, let the engagement estimation run on a local computer. The system shows an emoji representing the engagement state instead of the actual face. Only engagement-level data were sent in this case, and no actual affective data were sent. Although MeetmEE has not been provided with this feature, improvement in that direction is a potential measure.

Training/Workshop/User Manuals Authorities are encouraged to provide a user manual for the system and hold training regarding its use. When the human factor remains vulnerable to information security, humans must be influenced and trained to practice good judgment within a policy framework [103]. Therefore, a workshop is required to assist educators in using data to inform instruction. Besides technical training, ethical education through workshops is seen as a positive approach to avoid the risk of perpetuating existing patterns of unethical and inequitable technology development [209].

More importantly, the consent list, instruction, policy consent, and control mechanism must be made concise, clear, and transparent for users and understood by them [76]. After all, the effort towards information security must permeate all parties, both the organization and individual users.

5.4 Chapter Conclusion

This article discusses the design principle that was proposed to address the ethical gap in implementing automatic engagement estimation in distance learning practice. In addition to the user experience experiment results (Chapter 4), three ethical issues were considered when constructing the design principle to implement an automatic engagement estimation in distance learning practice.

The design principle (Figure 5.1) incorporates both technical and operational measures. Technical measures include the consideration of a trustworthy estimation model, architecture, and data security. Whereas the operational measures involve the participation of authorities such as education institutions, and the technology users, that is, educators and learners.

Chapter 6

Conclusion

This chapter contains the summary and highlights of this research contribution. Finally, the remaining works and ideas for improving this research in the future are also described.

6.1 Summary

This study addressed the practical knowledge gap of automatic recognition and analysis of learners' engagement in the distance learning process. In contrast to the other works that focus on ICT development, this study investigated automatic engagement estimation from the perspective of distance learning practices. This study mainly addresses the main research question, "How do educators or education institutions safely apply automatic engagement estimation in their distance learning process?", which is broken down into several research questions:

- **RQ1**: What requirements did the literature develop for automatic engagement estimation?
- **RQ2**: How to develop real-time engagement estimation tools for distance learning practice?
- **RQ3**: How to implement automatic engagement estimation in distance learning with taking into account distance learning characteristics and ethical impact?

To address the RQ1, we did a systematic review in Chapter 2 and experimented on LSTM and CNN models to build automatic engagement estimation in Chapter 3. From the engagement taxonomy, we introduced (Figure 2.2),

we can define what type of engagement is being measured from the available cues, the stimuli presented to the participant during data collection, and the observed physical or cognitive behaviours. Furthermore, the review of the dataset and machine learning-based methods used in the literature gives an insight into the current trend in automatic engagement estimation. Therefore, we believe that the combination of a clear engagement definition, a suitable dataset, and machine learning methods are the basic requirements to develop an automatic engagement estimation.

To address the RQ2, Chapter 3 technically investigated the requirements by developing automatic engagement estimation modules using LSTM and CNN, while Chapter 4 proposed system designs to implement the automatic engagement estimation module in distance learning practice.

As discussed in Chapter 2, LSTM is gaining popularity in engagement estimation research due to its sequential characteristics, while CNN is already the most popular method in computer vision-based engagement estimation. Likewise, the DAiSEE is the most popular and publicly available engagement dataset. Our engagement estimation models can successfully distinguish the engagement level. Although the prediction accuracy is low, we could analysed the suitability of the DAiSEE dataset and the feasibility of LSTM and CNN for real-time implementation. The experiment justifies the statement that incorrect interpretation and inconsistency in engagement measurement methods and annotations lead to severe bias. Moreover, LSTM is less feasible for practical implementation compared to CNN from a runtime perspective. However, we found that classic machine learning would be the best practice, especially for real-time engagement estimation. Finally, a framework for real-time automatic engagement estimation (Figure 3.10) was proposed for further implementation of automatic engagement estimation in distance learning practice.

Furthermore, we proposed design systems to develop a prototype for asynchronous and synchronous distance learning settings, i.e., RAMALAN, a real-time engagement assessment in an asynchronous system (Figure 4.4) based on the framework in Figure 3.10, and MeetmEE (Figure 4.7) for one-on-one synchronous distance learning.

A pilot experiment was conducted as a user experience survey to evaluate the MeetmEE system design and construct the ethical implementation design principle. The user experience experiment results show that most of the responses were very positive toward the automatic engagement estimation concept, represented in MeetmEE. Likewise, the user experience evaluation demonstrated positive perceptions of scales of stimulation, attractiveness, perspicuity, and novelty, yet low dependability and negative perception of efficiency due to low performance.

In addition to the user experience experiment results, three ethical issues were considered when constructing the design principle to implement an automatic engagement estimation in distance learning practice. The design principle incorporates both technical and operational measures. Technical measures include the consideration of a trustworthy estimation model, architecture, and data security. The operational measures involve the participation of authorities such as education institutions and technology users: educators and learners.

Finally, the combined proposed MeetmEE system design and the design principle for the implementation leads to the contribution of this work, i.e., an end-to-end integration of automatic engagement estimation.

6.2 Contribution

The main contributions of this study are as follows:

1. A systematic review of automatic engagement estimation: definition, datasets, and methods.

2. RAMALAN and MeetmEE system designs for asynchronous and synchronous distance learning.

3. Design principle of the automatic engagement estimation for ethical implementation.

6.3 Limitation

The knowledge and application of our model and method are meant to help educators and education institutions better understand learner engagement in distance learning settings with respect to privacy (e.g., learners' affective data and information). The ethical implementation mechanism aims to empower web and software developers to optimize privacy-enhancing distance learning technologies in this respect. However, further evaluation is required to justify this claim.

Furthermore, this study has several limitations as follows:

1. In the systematic review to address RQ1, 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 [175] or human-robot interactions [27]. 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 be applied in the education process.

2. We experimented with six pre-processing scenarios and found that Multilayer-LSTM with scenario 4 performs best. However, further discussion regarding the correlation between the pre-processing scenarios and the LSTM models is still open for future work.

3. The proposed MeetmEE system design currently only works for one-on-one meetings. Improving the scalability of the system design to enable multiple participants (e.g., an online classroom) would improve its practicality in synchronous distance learning practices.

4. Furthermore, the automatic engagement estimation model in MeetmEE merely defines the engagement level using frontal images and the distance of the participant from the monitor, which leads to the drawback of false interpretation. For example, some behaviours during the meeting, such as writing or doing other related assignments, where the face is directed away from the monitor, would be estimated as low engagement. Moreover, learners may look attentive while, in fact, not following at all [153]. Therefore, Including more engagement cues, such as log files and physiological cues, in addition to frontal-image analysis, will improve estimation outcome reliability.

5. Regarding output, our proposed system is useful for educators to know when learners lose their general engagement. However, a more detailed report is a challenge for future work, which not only provides individual feedback but also provides feedback to educators by putting them together. For example, implementing the prototype in an actual LMS is suggested for real implementation in the educational learning process.

6. Regarding the technical measures for ethical implementation based on the proposed design principle, the current MeetmEE system must be adjusted to meet the requirements, especially in terms of trustworthy data security. Furthermore, deploying a machine-learning model in practice, in this case, the end-to-end integration of the automatic engagement estimation model, remains challenging. Particularly in a synchronous learning session, in which audio-video stream data are the main input for the automatic engagement estimation module.

6.4 Future Work

Furthermore, we found several remaining challenges that have room for improvement, including MeetmEE scalability, cognitive engagement, personalized engagement, and machine-learning pitfalls.

MeetmEE Scalability MeetmEE will be more beneficial with bigger scalability, for instance, not only for one-on-one scenarios but also for multi-user. The system design can be improved by taking more consideration of software architecture and data communication between front-end applications, prediction servers, and back-end servers. Furthermore, the safe implementation can be improved with software measures, for instance, by adding an authentication login system and data anonymisation.

Cognitive Engagement Table 1 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 [126]. However, few studies (Table 1) have considered this type of engagement. For example, Turan et al. [210] have studied the relationship between facial expression and cognitive engagement. However, 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 [78]. 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 [177]. Several automatic engagement estimations target participants with specific cultures or backgrounds. For example, as shown by [128], 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 [175]. The existing engagement estimation research has used datasets based on one ethnicity [110]. Similarly, as we have done in Section 3, the participants in the DAiSEE dataset used in this study are of a single ethnicity. Engagement is an inner state that sometimes does not appear visually, which may be influenced by individual factors such as age, gender, ethnicity, and prior experience. Future automatic engagement estimation should consider individual factors and differences in the analysis for more reliable engagement estimation.

However, it is unknown how engagement estimation results can be generalized in actual applications [32]. 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) [205, 51]) to train the model [177]. 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. [182, 181]. 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 [201]. Although education research was not included in the review [181], 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 5-8). 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 [181]. Furthermore, education experts and ML experts could collaborate on engagement research to develop more effective models [201].

Bibliography

[1] Abdel-Nasser Sharkawy. Principle of Neural Network and Its Main Types: Review. *Journal of Advances in Applied & Computational Mathematics*, 7:8–19, 8 2020.

[2] Ali Abedi and Shehroz S. Khan. Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network. In *2021 18th Conference on Robots and Vision (CRV)*, pages 151–157, 2021.

[3] ACM International Conference on Multimodal Interaction 2020. Eighth Emotion Recognition in the Wild Challenge (EmotiW), 2020.

[4] Ahmad Qasim Mohammad AlHama. Students' data privacy: How far it is protected? (Ethical Percpective). In *2014 International Conference on Interactive Collaborative Learning (ICL)*, pages 619–622. IEEE, 12 2014.

[5] Rieks Akker, Dennis Hofs, Hendri Hondorp, Harm Akker, Job Zwiers, and Anton Nijholt. Supporting Engagement and Floor Control in Hybrid Meetings. pages 276–290. 2009.

[6] Soraia M. Alarcão and Manuel J. Fonseca. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2019.

[7] Karl L. Alexander, Doris R. Entwisle, and Carrie S. Horsey. From First Grade Forward: Early Foundations of High School Dropout. *Sociology of Education*, 70(2):87, 4 1997.

[8] Akhmad Alimudin and Aliv Faizal Muhammad. Online Video Conference System Using WebRTC Technology for Distance Learning Support. In *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, pages 384–387. IEEE, 10 2018.

[9] Khawlah Altuwairqi, Salma Kammoun Jarraya, Arwa Allinjawi, and Mohamed Hammami. A new emotion–based affective model to detect student's engagement. *Journal of King Saud University - Computer and Information Sciences*, 33(1):99–109, 1 2021.

[10] Khawlah Altuwairqi, Salma Kammoun Jarraya, Arwa Allinjawi, and Mohamed Hammami. Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image and Video Processing*, 15(7):1387–1395, 10 2021.

[11] Omar AlZoubi, Sidney K. D'Mello, and Rafael A. Calvo. Detecting Naturalistic Expressions of Nonbasic Affect Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(3):298–310, 7 2012.

[12] Mohd Anwar and Jim Greer. Enabling Reputation-Based Trust in Privacy-Enhanced Learning Systems. In *Intelligent Tutoring Systems*, pages 681–683. Springer Berlin Heidelberg, Berlin, Heidelberg.

[13] Andrea Apicella, Pasquale Arpaia, Mirco Frosolone, Giovanni Improta, Nicola Moccaldi, and Andrea Pollastro. EEG-based measurement system for monitoring student engagement in learning 4.0. *Scientific Reports*, 12(1):5857, 12 2022.

[14] T. S. Ashwin and Ram Mohana Reddy Guddeti. Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems*, 108:334–348, 7 2020.

[15] T. S. Ashwin and Ram Mohana Reddy Guddeti. 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, 11 2020.

[16] Roger Azevedo. Defining and Measuring Engagement and Learning in Science: Conceptual, Theoretical, Methodological, and Analytical Issues. *Educational Psychologist*, 50(1):84–94, 1 2015.

[17] Sileye O. Ba and Jean-Marc Odobez. Head Pose Tracking and Focus of Attention Recognition Algorithms in Meeting Rooms. In *Multimodal Technologies for Perception of Humans*, pages 345–357. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[18] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. 9 2014.

[19] Anna Kamille Balan, Adrian Rey Jacintos, and Thomas Montemayor. The Influence of Online Learning towards the Attention Span and Motivation of College Students. 2020.

[20] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015-Janua, 2015.

[21] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[22] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[23] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[24] Tobias Baur, Gregor Mehlmann, Ionut Damian, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin, Elisabeth André, and Patrick Gebhard. Context-Aware Automated Analysis and Annotation of Social Human–Agent Interactions. *ACM Transactions on Interactive Intelligent Systems*, 5(2):1–33, 7 2015.

[25] Atef Ben-Youssef, Chloe Clavel, and Slim Essid. Early Detection of User Engagement Breakdown in Spontaneous Human-Humanoid Interaction. *IEEE Transactions on Affective Computing*, 12(3):776–787, 7 2021.

[26] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 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*, pages 464–472, New York, NY, USA, 11 2017. ACM.

[27] Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 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, 12 2019.

[28] Yoshua Bengio. 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, pages 17–37, 2011.

[29] Dana Bevilacqua, Ido Davidesco, Lu Wan, Kim Chaloner, Jess Rowland, Mingzhou Ding, David Poeppel, and Suzanne Dikker. 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, 3 2019.

[30] Prakhar Bhardwaj, P. K. Gupta, Harsh Panwar, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, and Anubha Bhaik. Application of Deep Learning on Student Engagement in e-learning environments. *Computers and Electrical Engineering*, 93, 7 2021.

[31] Nigel Bosch. Detecting student engagement: Human versus machine. *UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 317–320, 2016.

[32] Nigel Bosch, Sidney K. D'mello, Jaclyn Ocumpaugh, Ryan S. Baker, and Valerie Shute. Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *ACM Transactions on Interactive Intelligent Systems*, 6(2):1–26, 8 2016.

[33] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 7 1997.

[34] Hennie Brugman and Albert Russel. Annotating Multi-media/Multimodal Resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 5 2004. European Language Resources Association (ELRA).

[35] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[36] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017-Janua, pages 1302–1310, 2017.

[37] Elena Carlotta Olivetti, Maria Grazia Violante, Enrico Vezzetti, Federica Marcolin, and Benoit Eynard. Engagement Evaluation in a Virtual Learning Environment via Facial Expression Recognition and Self-Reports: A Preliminary Approach. *Applied Sciences*, 10(1):314, 12 2019.

[38] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[39] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W. McOwan. 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*, pages 421–428, 2012.

[40] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497, 10 2019.

[41] Rebeca Cerezo, Miguel Sánchez-Santillán, M. Puerto Paule-Ruiz, and J. Carlos Núñez. Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, 96:42–54, 5 2016.

[42] Cheng Chang, Lei Chen, Cheng Zhang, and Yang Liu. An ensemble model using face and body tracking for engagement detection. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pages 616–622, 2018.

[43] Maher Chaouachi, Pierre Chalfoun, Imène Jraidi, and Claude Frasson. Affect and mental engagement: Towards adaptability for intelligent systems. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference, FLAIRS-23*, (Flairs):355–360, 2010.

[44] Iman Chatterjee, Maja Goršič, Joshua D. Clapp, and Domen Novak. Automatic Estimation of Interpersonal Engagement During Naturalistic Conversation Using Dyadic Physiological Measurements. *Frontiers in Neuroscience*, 15, 10 2021.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 6 2002.

[46] Xieling Chen, Haoran Xie, Di Zou, and Gwo Jen Hwang. Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1:100002, 1 2020.

[47] Yi-Wei Chen and Chih-Jen Lin. Combining SVMs with Various Feature Selection Strategies. In *Feature Extraction. Studies in Fuzziness and Soft Computing*, volume 207, pages 315–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[48] Michelene T. H. Chi and Ruth Wylie. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, 49(4):219–243, 10 2014.

[49] Sandra L. Christenson, Amy L. Reschly, and Cathy Wylie. *Handbook of Research on Student Engagement*. Springer US, Boston, MA, 2012.

[50] Mihaela Cocea and Stephan Weibelzahl. Disengagement detection in online learning: Validation studies and perspectives. *IEEE Transactions on Learning Technologies*, 4(2):114–124, 2011.

[51] Daniela Conti, Allegra Cattani, Santo Di Nuovo, and Alessandro Di Nuovo. 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)*, pages 555–560. IEEE, 8 2015.

[52] Diana K. Darnell and Paul A. Krieg. Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures. *PLOS ONE*, 14(12):e0225709, 12 2019.

[53] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1):20, 1979.

[54] Berardina De Carolis, Francesca D'Errico, Nicola Macchiarulo, and Giuseppe Palestra. "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*, pages 80–85, 2019.

[55] Dick de Ridder, David M J Tax, Bangjun Lei, Guangzhu Xu, Ming Feng, Yaobin Zou, and Ferdinand van der Heijden. *Classification, Parameter Estimation and State Estimation*. John Wiley & Sons, Ltd, Chichester, UK, 6 2017.

[56] Francesco Del Duchetto, Paul Baxter, and Marc Hanheide. Are You Still With Me? Continuous Engagement Assessment From a Robot's Point of View. *Frontiers in Robotics and AI*, 7, 9 2020.

[57] Kevin Delgado, Juan Manuel Origgi, Tania Hasanpoor, Hao Yu, Danielle Allessio, Ivon Arroyo, William Lee, Margrit Betke, Beverly Woolf, and Sarah Adel Bargal. Student Engagement Dataset. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2021-Octob, pages 3621–3629. Institute of Electrical and Electronics Engineers Inc., 2021.

[58] Didan Deng, Zhaokang Chen, Yuqian Zhou, and Bertram Shi. MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2621–2628, 4 2020.

[59] Jiankang Deng, J Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage Dense Face Localisation in the Wild. *ArXiv*, abs/1905.00641, 2019.

[60] M. Ali Akber Dewan, Fuhua Lin, Dunwei Wen, Mahbub Murshed, and Zia Uddin. 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)*, pages 1895–1902. IEEE, 10 2018.

[61] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *Proceedings of the 2018 on International Conference*

on Multimodal Interaction - ICMI '18, number October, pages 653–656, New York, New York, USA, 2018. ACM Press.

[62] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 784–789, New York, NY, USA, 10 2020. ACM.

[63] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 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, 9 2018.

[64] Sidney D'Mello, Ed Dieterle, and Angela Duckworth. Advanced, Analytic, Automated (AAA) Measurement of Engagement During Learning. *Educational Psychologist*, 52(2):104–123, 4 2017.

[65] Sidney D'Mello, Rosalind W. Picard, and Arthur Graesser. Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, 22(4):53–61, 7 2007.

[66] Ligeng Dong, Huijun Di, Linmi Tao, Guangyou Xu, and Patrick Oliver. Visual Focus of Attention Recognition in the Ambient Kitchen. pages 548–559. 2010.

[67] Denis Dresvyanskiy, Wolfgang Minker, and Alexey Karpov. Deep Learning Based Engagement Recognition in Highly Imbalanced Data. In *Speech and Computer*, pages 166–178, 2021.

[68] Ilana Dubovi. Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Computers & Education*, 183:104495, 7 2022.

[69] Yan (Catherine) Wu Eberhard Hechler, Maryela Weihrauch. *Data Fabric and Data Mesh Approaches with AI.* Apress Berkeley, CA, 4 2023.

[70] Naktal Moaid Edan, Ali Al-Sherbaz, and Scott Turner. Design and evaluation of browser-to-browser video conferencing in WebRTC. In *2017 Global Information Infrastructure and Networking Symposium (GIIS)*, pages 75–78. IEEE, 10 2017.

[71] Gudrun Eisele, Hugo Vachon, Ginette Lafit, Peter Kuppens, Marlies Houben, Inez Myin-Germeys, and Wolfgang Viechtbauer. 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, 3 2022.

[72] Paul Ekman and W.V. Friesen. Facial action coding system. *Palo Alto, CA: Consulting Psychologists Press*, 1978.

[73] Olov Engwall, Ronald Cumbal, José Lopes, Mikael Ljung, and Linnea Månsson. Identification of Low-engaged Learners in Robot-led Second Language Conversations with Adults. *ACM Transactions on Human-Robot Interaction*, 11(2):1–33, 6 2022.

[74] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, New York, NY, USA, 10 2013. ACM.

[75] Ali Farooq, Johanna Isoaho, Seppo Virtanen, and Jouni Isoaho. Information Security Awareness in Educational Institution: An Analysis of Students' Individual Factors. In *2015 IEEE Trustcom/BigDataSE/ISPA*, pages 352–359. IEEE, 8 2015.

[76] Denis Feth, Andreas Maier, and Svenja Polst. A User-Centered Model for Usable Security and Privacy. pages 74–89. 2017.

[77] Jeremy D. Finn and Kayla S. Zimmer. Student Engagement: What Is It? Why Does It Matter? In *Handbook of Research on Student Engagement*, pages 97–131. Springer US, Boston, MA, 2012.

[78] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, 74(1):59–109, 3 2004.

[79] Jennifer A. Fredricks and Wendy McColskey. The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments. In *Handbook of Research on Student Engagement*, pages 763–782. Springer US, Boston, MA, 2012.

[80] Md. Tahmid Hasan Fuad, Awal Ahmed Fime, Delowar Sikder, Md. Akil Raihan Iftee, Jakaria Rabbi, Mabrook S. Al-Rakhami, Abdu Gumaei, Ovishake Sen, Mohtasim Fuad, and Md. Nazrul Islam. Recent

Advances in Deep Learning Techniques for Face Recognition. *IEEE Access*, 9:99112–99142, 2021.

[81] B. Gabrys and A. Bargiela. General fuzzy min-max neural network for clustering and classification. *IEEE Transactions on Neural Networks*, 11(3):769–783, 5 2000.

[82] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. 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, 7 2012.

[83] V. García, J. S. Sánchez, and R. A. Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2 2012.

[84] Jeffrey M Girard. CARMA: Software for continuous affect rating and media annotation. *Journal of open research software*, 2(1), 2014.

[85] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review*, 33(1):27–49, 3 2021.

[86] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, volume MIT Press. MIT Press, 2016.

[87] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 4 2013.

[88] Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. Automatically recognizing facial expression: Predicting engagement and frustration. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*, 2013.

[89] Barbara A. Greene. Measuring Cognitive Engagement With Self-Report Scales: Reflections From Over 20 Years of Research. *Educational Psychologist*, 50(1):14–30, 1 2015.

[90] Amogh Gudi, H. Emrah Tasli, Tim M. den Uyl, and Andreas Maroulis. 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)*, volume 2015-Janua, pages 1–5, 2015.

[91] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. DAiSEE: Towards User Engagement Recognition in the Wild. 14(8):1–12, 9 2016.

[92] Mohammad Nehal Hasnine, Huyen T.T. Bui, Thuy Thi Thu Tran, Ho Tran Nguyen, Gškhan Akçapõnar, and Hiroshi Ueda. Students' emotion extraction and visualization for engagement detection in online learning. *Procedia Computer Science*, 192:3423–3431, 1 2021.

[93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[94] Javier Hernandez, Zicheng Liu, Geoff Hulten, Dave DeBarr, Kyle Krum, and Zhengyou Zhang. Measuring the engagement level of TV viewers. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2013.

[95] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 7 2006.

[96] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[97] G. Holmes, A. Donkin, and I.H. Witten. WEKA: a machine learning workbench. In *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, pages 357–361. IEEE.

[98] Tao Huang, Yunshan Mei, Hao Zhang, Sanya Liu, and Huali Yang. Fine-grained Engagement Recognition in Online Learning Environment. In *2019 IEEE 9th International Conference on Electronics Information*

and Emergency Communication (ICEIEC), pages 338–341. IEEE, 7 2019.

[99] Farzad Husain, Babette Dellen, and Carme Torras. Action Recognition Based on Efficient Deep Feature Learning in the Spatio-Temporal Domain. *IEEE Robotics and Automation Letters*, 1(2):984–991, 7 2016.

[100] Mushtaq Hussain, Wenhao Zhu, Wu Zhang, and Syed Muhammad Raza Abidi. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Computational Intelligence and Neuroscience*, 2018, 2018.

[101] Ian Witten and Eibe Frank. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, second edition, 6 2005.

[102] Ilya Grigorik. *High Performance Browser Networking: What Every Web Developer Should Know About Networking and Browser Performance*. O'Reilly Media, first edition, 9 2013.

[103] Damian Jackson and Paul Hayes. Ensuring Security of Data and Information Flow in Emergency Response Decision Support. In *2016 11th International Conference on Availability, Reliability and Security (ARES)*, pages 792–797. IEEE, 8 2016.

[104] Phichaya Jaturawat, Pasinee Pongmanawut, and Manop Phankokkruad. A remote image collecting to create initiative database with indexing and querying for enhance face recognition. In *2015 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 206–210. IEEE, 4 2015.

[105] Hunseop Jeong, Taehyung Lee, and Young Ik Eom. WebRTC-based Resource Offloading in Smart Home Environments. In *2022 IEEE International Conference on Consumer Electronics (ICCE)*, pages 01–06. IEEE, 1 2022.

[106] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 1 2013.

[107] Hideo Joho, Jacopo Staiano, Nicu Sebe, and Joemon M Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multim. Tools Appl.*, 51(2):505–523, 2011.

[108] Michael I Jordan. Attractor Dynamics and Parallelism in a Connectionist Sequential Machine. In *Artificial Neural Networks: Concept Learning*, pages 112–127. 1990.

[109] Shofiyati Nur Karimah and Shinobu Hasegawa. A Real-time Engagement Assessment in Online Learning Process Using Convolutional Neural Network. In *The 12th Asian Conference on Education (ACE2020)*, pages 437–448, 1 2020.

[110] Shofiyati Nur Karimah and Shinobu Hasegawa. Automatic Engagement Recognition for Distance Learning Systems: A Literature Study of Engagement Datasets and Methods. In *International Conference on Human-Computer Interaction*, pages 264–276, 2021.

[111] Shofiyati Nur Karimah and Shinobu Hasegawa. Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 9(1):31, 11 2022.

[112] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and Localization of Student Engagement in the Wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 12 2018.

[113] Deb Keen. Engagement of Children With Autism in Learning. *Australasian Journal of Special Education*, 33(2):130–140, 10 2009.

[114] Kenneth C. Laudon and Jane P. Laudon. *Management Information Systems: Managing the Digital Firm*. Pearson Education Limited, 13 edition, 2014.

[115] Michael Kipp. Spatiotemporal Coding in ANVIL. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 5 2008. European Language Resources Association (ELRA).

[116] Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35, 11 2018.

[117] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 5 2017.

[118] Hennie Kruger, Lynette Drevin, and Tjaart Steyn. A vocabulary test to assess information security awareness. *Information Management & Computer Security*, 18(5):316–327, 11 2010.

[119] Justin Kruger and David Dunning. 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, 1999.

[120] Mikel Labayen, Ricardo Vea, Julian Florez, Naiara Aginako, and Basilio Sierra. Online Student Authentication and Proctoring System Based on Multimodal Biometrics Technology. *IEEE Access*, 9:72398–72411, 2021.

[121] Reed Larson and Mihaly Csikszentmihalyi. The Experience Sampling Method. In *Flow and the Foundations of Positive Psychology*, pages 21–34. Springer Netherlands, Dordrecht, 2014.

[122] Hao Lei, Yunhuo Cui, and Wenye Zhou. Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46(3):517–528, 3 2018.

[123] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. Comparing Models of Disengagement in Individual and Group Interactions. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 99–105, New York, NY, USA, 3 2015. ACM.

[124] Shan Li and Weihong Deng. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, 3045(c):1–1, 2020.

[125] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 7 2017.

[126] Shan Li, Susanne P. Lajoie, Juan Zheng, Hongbin Wu, and Huaqin Cheng. Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education*, 163:104114, 4 2021.

[127] Jiacheng Liao, Yan Liang, and Jiahui Pan. Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, 51(10):6609–6621, 10 2021.

[128] A.V. Libin and E.V. Libin. Person-robot interactions from the robopsychologists' point of view: the robotic psychology and robotherapy approach. *Proceedings of the IEEE*, 92(11):1789–1803, 11 2004.

[129] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 10 2017.

[130] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (CERT). In *Face and Gesture 2011*, pages 298–305. IEEE, 3 2011.

[131] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning Expressionlets on Spatio-temporal Manifold for Dynamic Facial Expression Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[132] Salvatore Loreto and Simon Pietro Romano. Real-Time Communications in the Web: Issues, Achievements, and Ongoing Standardization Efforts. *IEEE Internet Computing*, 16(5):68–73, 9 2012.

[133] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Face and Gesture 2011*, pages 57–64. IEEE, 3 2011.

[134] Dubi Lufi and Iris Haimov. Effects of age on attention level: changes in performance between the ages of 12 and 90. *Aging, Neuropsychology, and Cognition*, 26(6):904–919, 11 2019.

[135] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with Gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205. IEEE Comput. Soc, 2002.

[136] Xiaoyang Ma, Min Xu, Yao Dong, and Zhong Sun. Automatic student engagement in online learning environment based on neural turing machine. *International Journal of Information and Education Technology*, 11(3):107–111, 3 2021.

[137] Jayawant N. Mandrekar. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 9 2010.

[138] Simon J. Mason and Andreas P. Weigel. A Generic Forecast Verification Framework for Administrative Purposes. *Monthly Weather Review*, 137(1):331–349, 1 2009.

[139] Rob McCarney, James Warner, Steve Iliffe, Robbert van Haselen, Mark Griffin, and Peter Fisher. The Hawthorne Effect: a randomised, controlled trial. *BMC Medical Research Methodology*, 7(1):30, 12 2007.

[140] Karen S. McNeal, Min Zhong, Nick A. Soltis, Lindsay Doukopoulos, Elijah T. Johnson, Stephanie Courtney, Akilah Alwan, and Mallory Porch. Biosensors Show Promise as a Measure of Student Engagement in a Large Introductory Biology Course. *CBE—Life Sciences Education*, 19(4):ar50, 12 2020.

[141] Naval Kishore Mehta, Shyam Sunder Prasad, Sumeet Saurav, Ravi Saini, and Sanjay Singh. Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*, 3 2022.

[142] Matthew Militello, Lisa Bass, Karen Jackson, and Yuling Wang. How Data Are Used and Misused in Schools: Perceptions from Teachers and Principals. *Education Sciences*, 3(2):98–120, 4 2013.

[143] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. volume 2, pages 273–289. 8 2020.

[144] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 1 2019.

[145] Torin Monahan and Jill A. Fisher. Benefits of 'observer effects': lessons from the field. *Qualitative Research*, 10(3):357–376, 6 2010.

[146] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 1 2017.

[147] Mahbub Murshed, M. Ali Akber Dewan, Fuhua Lin, and Dunwei Wen. Engagement Detection in e-Learning Environments using Convolutional Neural Networks. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 80–86, 2019.

[148] Omid Mohamad Nezami, Deborah Richards, and Len Hamey. Semi-supervised detection of student engagement. In *Proceedings ot the 21st Pacific Asia Conference on Information Systems: "Societal Transformation Through IS/IT", PACIS 2017*, 2017.

[149] Manuel Ninaus, Simon Greipl, Kristian Kiili, Antero Lindstedt, Stefan Huber, Elise Klein, Hans-Otto Karnath, and Korbinian Moeller. Increased emotional engagement in game-based learning – A machine learning approach on facial emotion detection data. *Computers & Education*, 142:103641, 12 2019.

[150] Xuesong Niu, Hu Han, Jiabei Zeng, Xuran Sun, Shiguang Shan, Yan Huang, Songfan Yang, and Xilin Chen. Automatic Engagement Prediction with GAP Feature. In *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, number 1, pages 599–603, New York, New York, USA, 2018. ACM Press.

[151] Thiago Nóbrega, Carlos Eduardo S. Pires, and Dimas Cassimiro Nascimento. Blockchain-based Privacy-Preserving Record Linkage: enhancing data privacy in an untrusted environment. *Information Systems*, 102:101826, 12 2021.

[152] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015.

[153] Barbara Oakley, Beth Rogowsky, and Terrence J. Sejnowski. *Uncommon Sense Teaching: Practical Insights in Brain Science to Help Students Learn*. 2021.

[154] Heather L. O'Brien and Elaine G. Toms. 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, 1 2010.

[155] F. Okubo, T. Yamashita, A. Shimada, and H. Ogata. A neural network approach for students' performance prediction. In *Proceedings of the*

Seventh International Learning Analytics & Knowledge Conference, pages 598–599, New York, NY, USA, 3 2017. ACM.

[156] Fan Ouyang and Pengcheng Jiao. Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2:100020, 1 2021.

[157] Chakradhar Pabba and Praveen Kumar. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, 39(1), 1 2022.

[158] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1):89, 12 2021.

[159] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. In *Procedings of the British Machine Vision Conference 2015*, pages 1–12, 2015.

[160] Gordon Pennycook, Robert M. Ross, Derek J. Koehler, and Jonathan A. Fugelsang. Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, 24(6):1774–1784, 12 2017.

[161] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. A Model of Attention and Interest Using Gaze Behavior. pages 229–240. 2005.

[162] Penelope L. Peterson, Susan R. Swing, Kevin D. Stark, and Gregory A. Waas. Students' Cognitions and Time on Task during Mathematics Instruction. *American Educational Research Journal*, 21(3):487–515, 1984.

[163] Manop Phankokkruad and Phichaya Jaturawat. An evaluation of technical study and performance for real-time face detection using Web Real-Time Communication. In *2015 International Conference on Computer, Communications, and Control Technology (I4CT)*, pages 162–166. IEEE, 4 2015.

[164] Claire Cameron Ponitz, Sara E. Rimm-Kaufman, Kevin J. Grimm, and Timothy W. Curby. Kindergarten Classroom Quality, Behavioral Engagement, and Reading Achievement. *School Psychology Review*, 38(1):102–120, 3 2009.

[165] Andreas Trier Poulsen, Simon Kamronn, Jacek Dmochowski, Lucas C. Parra, and Lars Kai Hansen. EEG in the classroom: Synchronised neural recordings during video presentation. *Scientific Reports*, 7(1):43916, 4 2017.

[166] Athanasios Psaltis, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Games*, 10(3):292–303, 2018.

[167] Athanasios Psaltis, Kyriaki Kaza, Kiriakos Stefanidis, Spyridon Thermos, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal affective state recognition in serious games applications. *IST 2016 - 2016 IEEE International Conference on Imaging Systems and Techniques, Proceedings*, pages 435–439, 2016.

[168] Weizheng Qiao and Xiaojun Bi. Ternary-task convolutional bidirectional neural turing machine for assessment of EEG-based cognitive workload. *Biomedical Signal Processing and Control*, 57:101745, 3 2020.

[169] Vikram Ramanarayanan, Chee Wee Leong, and David Suendermann-Oeft. Rushing to Judgement: How do Laypeople Rate Caller Engagement in Thin-Slice Videos of Human–Machine Dialog? In *Interspeech 2017*, pages 2526–2530, ISCA, 8 2017. ISCA.

[170] Vikram Ramanarayanan, Chee Wee Leong, David Suendermann-Oeft, and Keelan Evanini. 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*, pages 281–287, New York, NY, USA, 11 2017. ACM.

[171] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 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, pages 91–99, Cambridge, MA, USA, 2015. MIT Press.

[172] Fernando Ribeiro Trindade and Deller James Ferreira. Student Performance Prediction Based on a Framework of Teacher's Features. *Inter-*

national Journal for Innovation Education and Research, 9(2):178–196, 2 2021.

[173] Simon Pietro Romano, Salvatore Loreto, and Carol Davids. Real Time Communications in the Web: Current Achievements and Future Perspectives. *IEEE Communications Standards Magazine*, 1(2):20–21, 2017.

[174] Philipp V. Rouast, Marc T. P. Adam, and Raymond Chiong. Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 4 2021.

[175] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19), 6 2018.

[176] Ognjen Rudovic, Hae Won Park, John Busche, Bjorn Schuller, Cynthia Breazeal, and Rosalind W. Picard. 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)*, pages 217–226, 2019.

[177] Ognjen Rudovic, Yuria Utsumi, Jaeryoung Lee, Javier Hernandez, Eduardo Castelló Ferrer, Björn Schuller, and Rosalind W. Picard. 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*, pages 339–346, 2018.

[178] Ognjen Rudovic, Meiru Zhang, Bjorn Schuller, and Rosalind Picard. Multi-modal Active Learning From Human Data: A Deep Reinforcement Learning Approach. In *2019 International Conference on Multimodal Interaction*, pages 6–15, New York, NY, USA, 10 2019. ACM.

[179] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[180] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, pages 305–311, 2011.

[181] Sayash Kapoor and Arvind Narayanan. Leakage and the Reproducibility Crisis in ML-based Science. 7 2022.

[182] Sayash Kapoor, Priyanka Nanayakkara, Kenny Peng, Hien Pham, and Arvind Narayanan. The Reproducibility Crisis in ML-based Science, 2022.

[183] Bart W. Schermer, Bart Custers, and Simone van der Hof. The crisis of consent: how stronger legal protection may lead to weaker consent in data protection. *Ethics and Information Technology*, 3 2014.

[184] Gianluca Schiavo, Alessandro Cappelletti, and Massimo Zancanaro. Engagement recognition using easily detectable behavioral cues. *Intelligenza Artificiale*, 8(2):197–210, 2014.

[185] Adam Schmidt and Andrzej Kasiński. The Performance of the Haar Cascade Classifiers Applied to the Face and Eyes Detection. pages 816–823. 2007.

[186] Michael F. Schober and Herbert H. Clark. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232, 4 1989.

[187] Martin Schrepp. User experience questionnaire handbook, 2023. Accessed: 2023-11-19.

[188] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[189] Björn Schuller. Deep Learning Our Everyday Emotions. pages 339–346. 2015.

[190] Renee Schulz, Ghislain Maurice Isabwe, and Frank Reichert. Ethical issues of gamified ICT tools for higher education. In *2015 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)*, pages 27–31. IEEE, 8 2015.

[191] Renee Schulz, Ghislain Maurice Isabwe, and Frank Reichert. Investigating teachers motivation to use ICT tools in higher education. In *2015 Internet Technologies and Applications (ITA)*, pages 62–67, 2015.

[192] Shabnam Abtahi, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariri. YawDD: Yawning Detection Dataset, 2020.

[193] Junge Shen, Haopeng Yang, Jiawei Li, and Zhiyong Cheng. Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimedia Systems*, 28(2):469–478, 4 2022.

[194] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–14, 9 2014.

[195] P.K. Simpson. Fuzzy min-max neural networks. I. Classification. *IEEE Transactions on Neural Networks*, 3(5):776–786, 1992.

[196] George Suciu, Stefan Stefanescu, Cristian Beceanu, and Marian Ceaparu. WebRTC role in real-time communication and video conferencing. In *2020 Global Internet of Things Summit (GIoTS)*, pages 1–6. IEEE, 6 2020.

[197] Omer Sumer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal Engagement Analysis from Facial Videos in the Classroom. *IEEE Transactions on Affective Computing*, 2021.

[198] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[199] Ashwin T. S and Ram Mohana Reddy Guddeti. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, 25(2):1387–1415, 2020.

[200] Tadas Baltrusaitis. Output Format, 10 2019.

[201] The Batch team. Bad Machine Learning Makes Bad Science, 8 2022.

[202] Michael Thiruthuvanathan, Balachandran Krishnan, and M A Dorai Rangaswamy. Engagement Detection through Facial Emotional Recognition Using a Shallow Residual Convolutional Neural Networks. *International Journal of Intelligent Engineering and Systems*, 14:236–247, 2021.

[203] Chinchu Thomas, K.A.V. Puneeth Sarma, Srujan Swaroop Gajula, and Dinesh Babu Jayagopi. Automatic prediction of presentation style and

student engagement from videos. *Computers and Education: Artificial Intelligence*, 3:100079, 2022.

[204] Van Thong Huynh, Soo-Hyung Kim, Guee-Sang Lee, and Hyung-Jeong Yang. Engagement Intensity Prediction withFacial Behavior Features. In *2019 International Conference on Multimodal Interaction*, pages 567–571, New York, NY, USA, 10 2019. ACM.

[205] Matt Tincani, Jason Travers, and Amanda Boutot. 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, 9 2009.

[206] Tingfan Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett, and J. R. Movellan. Multilayer Architectures for Facial Action Unit Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1027–1038, 8 2012.

[207] Tom L. Beauchamp and James F. Childress. *Principles of biomedical ethics*. Oxford University Press, New York, 5 edition, 2001.

[208] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, volume 2015 Inter, pages 4489–4497, 2015.

[209] Conor Truax, Alexi Orchard, and Heather A. Love. The influence of curriculum and internship culture on developing ethical technologists: A case study of the University of Waterloo. In *2021 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–8. IEEE, 10 2021.

[210] Cigdem Turan, Karl David Neergaard, and Kin-Man Lam. Facial Expressions of Comprehension (FEC). *IEEE Transactions on Affective Computing*, 13(1):335–346, 1 2022.

[211] Yury A. Ushakov, Margarita V. Ushakova, Alexander E. Shukhman, Petr N. Polezhaev, and Leonid V. Legashev. WebRTC based Platform for Video Conferencing in An Educational Environment. In *2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE, 10 2019.

[212] Pieter Vanneste, José Oramas, Thomas Verelst, Tinne Tuytelaars, Annelies Raes, Fien Depaepe, and Wim Van den Noortgate. Computer

vision and human behaviour, emotion and cognition detection: A use case on student engagement. *Mathematics*, 9(3):1–20, 2 2021.

[213] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[214] P. Viola and M. Jones. 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*, volume 1, pages I–511–I–518, 2001.

[215] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 5 2004.

[216] Michael Voit and Rainer Stiefelhagen. 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*, page 173, New York, New York, USA, 2008. ACM Press.

[217] J. Wagner, Jonghwa Kim, and E. Andre. From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In *2005 IEEE International Conference on Multimedia and Expo*, pages 940–943. IEEE.

[218] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 3 2021.

[219] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 11 2010.

[220] Yuehua Wang, Anuhya Kotha, Pei Heng Hong, and Meikang Qiu. 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*, pages 270–275, 2020.

[221] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[222] Matthias Wenzel and Christoph Meinel. Full-body WebRTC video conferencing in a web-based real-time collaboration system. In *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 334–339. IEEE, 5 2016.

[223] Jacob Whitehill, Zewelanji Serpell, Yi Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[224] Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 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)*, pages 6204–6208, 2018.

[225] Philip H. Winne and Nancy E. Perry. Measuring Self-Regulated Learning. *Handbook of Self-Regulation*, pages 531–566, 1 2000.

[226] Peter Wittenburg, Hennie Brugman, Albert Russel, Alexander Klassmann, and Han Sloetjes. ELAN: a Professional Framework for Multimodality Research. In *LREC*, 2006.

[227] Christopher A. Wolters and Daniel J. Taylor. A Self-regulated Learning Perspective on Student Engagement. In *Handbook of Research on Student Engagement*, pages 635–651. Springer US, Boston, MA, 2012.

[228] Erroll Wood, Tadas Baltruaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, volume 2015 Inter, pages 3756–3764, 2015.

[229] Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3/4):129, 2009.

[230] Jianming Wu, Bo Yang, Yanan Wang, and Gen Hattori. 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*, pages 777–783, New York, NY, USA, 10 2020. ACM.

[231] Kui Xie, Benjamin C. Heddy, and Barbara A. Greene. Affordances of using mobile technology to support experience-sampling method in examining college students' engagement. *Computers & Education*, 128:183–198, 1 2019.

[232] D. Yang, Abeer Alsadoon, P. W.C. Prasad, A. K. Singh, and A. Elchouemi. An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment. In *Procedia Computer Science*, volume 125, pages 2–10. Elsevier B.V., 2018.

[233] Ji Won You. Identifying significant indicators using LMS data to predict course achievement in online learning. *The Internet and Higher Education*, 29:23–30, 4 2016.

[234] Jia Yue, Feng Tian, Kuo-Min Chao, Nazaraf Shah, Longzhuang Li, Yan Chen, and Qinghua Zheng. Recognizing Multidimensional Engagement of E-Learners Based on Multi-Channel Data in E-Learning Environment. *IEEE Access*, 7:149554–149567, 2019.

[235] Sang-Seok Yun, Mun-Taek Choi, Munsang Kim, and Jae-Bok Song. Intention Reading from a Fuzzy-Based Human Engagement Model and Behavioural Features. *International Journal of Advanced Robotic Systems*, 9(2), 8 2012.

[236] Woo-Han Yun, Dongjin Lee, Chankyu Park, and Jaehong Kim. Automatic Engagement Level Estimation of Kids in a Learning Environment. *International Journal of Machine Learning and Computing*, 5(2):148–152, 4 2015.

[237] Woo Han Yun, Dongjin Lee, Chankyu Park, Jaehong Kim, and Junmo Kim. Automatic Recognition of Children Engagement from Facial Video Using Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, 11(4):696–707, 10 2020.

[238] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, volume 2018-Janua, pages 2519–2528, 2017.

[239] Janez Zaletelj and Andrej Košir. Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing*, 2017(1):80, 12 2017.

[240] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Transactions on Affective Computing*, 8(3):300–313, 7 2017.

[241] Hao Zhang, Xiaofan Xiao, Tao Huang, Sanya Liu, Yu Xia, and Jia Li. An Novel End-to-end Network for Automatic Student Engagement Recognition. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 342–345. IEEE, 7 2019.

[242] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 10 2016.

[243] Zhaoli Zhang, Zhenhua Li, Hai Liu, Taihe Cao, and Sannyuya Liu. Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology. *Journal of Educational Computing Research*, 58(1):63–86, 3 2020.

[244] Zhenqiu Zhang, Yuxiao Hu, Ming Liu, and Thomas Huang. Head Pose Estimation in Seminar Room Using Multi View Face Detectors. pages 299–304. 2007.

[245] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. Affective Computing for Large-scale Heterogeneous Multimedia Data. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15(3s):1–32, 11 2019.

[246] Xianwen Zheng, Shinobu Hasegawa, Minh-Tuan Tran, Koichi Ota, and Teruhiko Unoki. Estimation of Learners' Engagement Using Face and Body Features by Transfer Learning. pages 541–552. 2021.

[247] Bin Zhu, Xinjie Lan, Xin Guo, Kenneth E. Barner, and Charles Boncelet. Multi-rate Attention Based GRU Model for Engagement Prediction. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 841–848, New York, NY, USA, 10 2020. ACM.

[248] Tianqing Zhu, Jin Li, Xiangyu Hu, Ping Xiong, and Wanlei Zhou. The Dynamic Privacy-Preserving Mechanisms for Online Dynamic Social

Networks. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2962–2974, 6 2022.

[249] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face Alignment Across Large Poses: A 3D Solution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-Decem, pages 146–155, 2016.

[250] Michael Zimmer. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325, 12 2010.

[251] Khalid Ibn Zinnah Apu, Nafiz Mahmud, Firoz Hasan, and Sabbir Hossain Sagar. P2P video conferencing system based on WebRTC. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 557–561. IEEE, 2 2017.

Appendix A – Systematic Review Method

A.1 – PRISMA Phase

A.1.1 - 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 automatic 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.

A.1.2 - 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.

A.1.3 - 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.

A.1.4 - 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.

A.2 – Literature Tables

A.2.1 - Engagement defined in the selected articles.

Table 1: An overview of the selected articles to address RQ1.

	Domains			Engagement Type				Engagement Cues			
Author	HCI	HRI	A/H	Cls.	B.	E.	C.	Aff.	BE	PC	LF
Wang et al. [219]	●					○		●			
Cocea et al. [50]	●				○						●
AlZoubi et al. [11]			A			○				●	
S-Syun et al. [235]		●				○		●			
Whitehill et al. [223]	●					○		●			
Schiavo et al. [184]	●				○	○		●			
W-H Yun et al. [236]	●					○		●			
Gupta et al. [91]	●					○		●			
Zaletelj et al. [239]				●	○			●			
Monkaresi [146]	●					○		●		●	
Youssef et al. [26]		●									
Zhalehpour et al. [240]	●										
Hussain et al. [100]	●				○		○				●
Psaltis et al. [166]	●					○		●	●		
Rudovic et al. [175]		●			○			●		●	
Ninaus et al. [149]	●					○			●		
Yue et al. [234]		●			○	○	○	●	●		●
Mollahosseini et al. [144]	●					○		●			
Celiktutan [40]		●			○			●		●	
Youssef et al. [27]		●				○		●			
Olivetti et al. [37]	●					○			●		
Ashwin et al. [199]				●	○	○		●			
Ashwin et al. [14]				●		○		●	●		
Pabba et al. [157]				●		○		●	●		
Duchetto [56]		●				○		●			
Yun et al. [237]		●			○	○		●			
Zhang et al. [243]	●				○	○		●			●
Liao et al. [127]	●					○		●			
Li et al. [126]	●					○	○	●			●
Bhardwaj [30]	●					○		●	●		
Goldberg [85]				●	○	○	○	●			
Chatterjee [44]			H			○				●	
Youssef et al. [25]		●			○			●			
Sümer et al. [197]				●		○		●			
Trindade [172]	●						○				●
Ma et al. ([136])	●					○		●			
Thiruthvanathan et al. [202]	●					○		●			
Altuwairqi [10]	●				○	○		●			●
Vanneste [212]				●	○	○		●			
Hasnine et al. [92]	●					○			●		
Delgado et al. ([57])	●					○		●			
Engwall et al. [73]		●				○		●	●		
Mehta et al. [141]	●					○		●			
Dubovi et al. [68]	●					○	○	●		●	
Thomas et al. [203]	●					○		●			●
Shen et al. [193]	●					○		●	●		
Apicella et al. [13]	●					○	○			●	

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).

A.2.2 - Engagement dataset

Non-Public engagement datasets (Table 2, Table 3) and engagement-related dataset (Table 4).

Table 2: Non-Public engagement dataset.

Dataset	Setting	Stimuli	Participants	Samples	Annotators	Label
Cocea et al. [50]	WE	An online course (HTML-tutor) in 7 sessions	48 users	14 logged events	3 EA	engaged, disengaged, or neutral
AlZoubi et al. [11]	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.
Whitehill et al. ([223])	S	A cognitive skills training software.	34 undergraduate students	10 seconds videos	7EA	Engage, Not Engage ([Very engaged,Engaged],[Nominally engaged,Not Engaged])
Schiavo et al. ([184])	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 Experience-Sampling Method (ESM) [121]	Neutral, **Engaged**, Stress
Woo-Han Yun et al. ([236])	S	An interactive testing software.	12 Children	2,745 of 30-second video clips	1 EA	4 engagement levels: high/low interest, low/high boredom
Zaletelj et al. ([239])	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)
Monkaresi et al. ([146])	S	Writing task (draft-feedback-review)	22 students	1,325 video segments	Concurrent and retrospective self-report	Not engaged, engaged
Hussain et al. ([100])	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 a total number of clicks - average clicks), then label = high engagement. Otherwise -¿ Low engagement
Rudovic et al. ([175])	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
Olivetti et al. ([37])	S	A virtual learning environment (A European Entrepreneurship VET Model and Assessment)	12 participants (6 F, 6 M)	3D videos	2 EA and self-report	Engagement level 1,2,3
Ashwin et al. ([199])	S&P	Offline classroom	50 students	24000 posed images of 50 students, 36000 images spontaneous	self-annotate and 2 EA	**Engaged**, boredom and neutral

S - Spontaneous; **P** - Posed; **W** - in-the-Wild; **WE** - Web-based learning environment; **W** - in-the-Wild; **EA** - External annotator;

Table 3: (Cont. 1) non-public engagement dataset.

Dataset	Setting	Stimuli	Participants	Samples	Annotators	Label
Ashwin et al. ([14])	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).
Pabba et al. ([157])	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)
Duchetto et al. ([56])	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
Yun et al. ([237])	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]
Zhang et al. ([243])			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 classifications Engaged, not engaged)
Goldberg et al. ([85])	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. ([44])	S	Dyadic conversation	16 dyads	Naturalistic conversations (15 minutes)	self-report	Engagement level (none to very high), Engagement scale (0-100)
Youssef et al. ([25])	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 a single user)	EA	Signs of User Engagement Breakdown (UEB): Breakdown, No Breakdown
Sumer et al. ([197])	S	Offline classroom	15 students	360 audio-visual recording	2 EA using CARMA every second	3 engagement class label (0,1,2)
Altuwairqi 2021 et al. ([10])	S	Writing task	42 participants	164 videos, mouse and keyboard log	self-annotation	strong, high, and medium engagements
Vanneste et al. ([212])	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
Hasnine et al. ([92])	S	interactive lecture, lecture video taken from YouTube (28s)	11 students		N/A (concentration index (CI))	Highly-engaged, engaged, disengaged
Delgado et al. ([57])	WE	Math problem on MathSpring.org	19 students	400 videos(18,721 frames)	3 EA	**Engaged** (looking at the screen or looking at their paper), wandering
Engwall et al. ([73])	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 disengaged, disengaged, neutral, engaged, very engaged)
Apicella et al. ([13])	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

S - Spontaneous; **P** - Posed; **W** - in-the-Wild; **WE** - Web-based learning environment; **W** - in-the-Wild; **EA** - External annotator;

Table 4: (Cont. 2) non-Public engagement-related dataset.

Dataset	Setting	Stimuli	Participants	Samples	Annotators	Label
Ninaus et al. ([149])	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"
Yue et al. ([234])	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
Li et al. ([126])	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)
Trindade et al. ([172])	WE	Courses in Moodle		2752 Moodle record data from 2015-2019		
Dubovi et al. ([68])	S	1. The Medication Administration Test (MAT), 3 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)[221]

S - Spontaneous; **P** - Posed; **W** - in-the-Wild; **WE** - Web-based learning environment; **W** - in-the-Wild; **EA** - External annotator;

A.2.2 - Machine Learning-based Methods

Table 5: 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. [219]	Thermal camera	Grayscale images pixels	**Feature extraction**: PCA, PCA + LDA, AAM, and AAM+LDA. **Classification**: KNN.**Validation**: LOOCV	Accuracy
Cocea et al. [50]	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) Attribute 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. [11]	3 sensors (electrocardiogram (ECG), facial electromyogram (EMG), galvanic skin response (GSR)), webcam, screen recorder.	117 features (EEG, corrugator muscle EMG, finger tips GSR)	**Preprocess**: low/high pass filter. **Feature extraction**: using Augsburg Biosignal Toolbox [217].**Classification**: PRTools 4.0 [55], a pattern recognition library for Matlab. 9 classifiers: 1) SVM with the 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. [235]	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%
Whitehill et al. [223]	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. [184]	Camera	Head movement and face features	**Features extraction**: using face actions and expression recognition [107]. **3-class classification**: SVM. **Validation**: LOOCV	Accuracy=73%, F-score = 63%
Woo-Han Yun et al. [236]	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. [91]	Camera	Image pixels	**Classification**: InceptionNet, C3D, LRCN.	Accuracy
Zaletelj et al. [239]	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. [146]	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.
Zhalehpour et al. [240]	Camera	Images	**Face tracking**: CHEHRA tracker. **Classification**: SVM. Accuracy: 5-class classification = 75.32%, 8-class = 65.84%	
Hussain et al. [100]	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
Psaltis et al. [166]	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%

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

Table 6: (Cont. 1) An overview of the method used in the selected articles to address RQ3.

Author	Input Device/Modality	Input Features	Estimation Method	Performance Metrics
Rudovic et al. [175]	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 [74], 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% ± 24 (average ± SD)
Ninaus et al. [149]	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 ≈ 64.18%
Yue et al. [234]	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. [144]	N/A	Images	CNN (AlexNet) and SVR on Valnce and Arousal labels	RMSE, CORR, SAGR, CCC.
Celiktutan et al. ([40])	Cameras (2 static & 2 dynamic), 2 biosensors	Image, sensor data	**Binary classifictions**: SVMs. **Validation**: a double LOOCV.	
Youssef et al. [27]	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. [37]	Camera	Images (geometrical description)	**3-class classification**: SVM	The classification result was compared with the questionnaire.
Ashwin et al. [199]	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 a single student in a 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. [14]	Camera	Images with facial expressions, hand gestures and body postures present	**Classification**: CNN with pre-trained on GoogleNet architecture[117]. **Validation**: 10-fold cross-validation.	Accuracy = 76%
Pabba et al. [157]	Camera	48x48 image pixels	Add additional public dataset: BAUM-1,DAiSEE, and Yawning Detection Dataset (YawDD)[192]. **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. [56]	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. [237]	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. [243]	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).	

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

Table 7: (Cont. 2) An overview of the method used in the selected articles to address RQ3.

Author	Input Device/Modality	Input Features	Estimation Method	Performance Metrics
Liao et al. [127]	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) = pre-trained 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. [126]	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 feature. **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. [30]	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. [85]	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 correlation coefficient between manual annotations' mean level and prediction models $r = .70$, $p = 0$
Chatterjee et al. [44]	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, least-squares boosting, and random forest implemented in MATLAB 2020b. **Validation**: LOOCV	
Youssef et al. [25]	Robot's camera	Distance; head, gaze and face streams; Speech; Laser	Face detection: NAOqi People Perception.**Face extraction**: OKAO Vision software. Imbalanced issue: undersampling "No breakdown", oversampling "Breakdown" class using SMOTE. **Binary classification**: LR, LDA, RF, and MLP. **Validation**: 5-fold cross-validation.	AUC ≈ 0.72
Sümer et al. [197]	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 is better than Affect, given that the criteria for the manual annotation of engagement are not directly related to gaze direction or facial expression.
Trindade et al. [172]	Log file	Teacher and students attributes	**WEKA**. Random Forest generated the best result.	AUC
Ma et al. [136]	Use DAiSEE	Eye gaze, facial action unit, head pose (117 dimensions); and body pose (60 dimensions)	**Feature extraction**: OpenFace 2.0. **Preprocess**: 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. ([202])	Indian origin faces datasets DAiSEE, iSAFE, ISED	508 images from ISED and iSAFE, 5295 images from DAiSEE.	**Feature extraction**: lightweight ResNet. **Classification**: ResNet classifier (CNN with 50 layers deep).	Accuracy, Precision, Recall, Sensitivity, Specificity, and F1 score
Altuwairqi 2021 et al. ([10])	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.
Vanneste et al. ([212])	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) [38]. **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.

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

Table 8: (Cont. 3) An overview of the method used in the selected articles to address RQ3.

Author	Input Device/Modality	Input Features	Estimation Method	Performance Metrics
Hasnine et al. ([92])	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. ([57])	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. ([73])	Cameras and microphone	Audio and visual features	**Feature extraction**: OpenFace 2.0. **Feature selection**: verbal classifications using bag-of-words representations, acoustic-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 happiness = High, neutral = Neutral, boredom and sadness = Low. **Engagement classification through face expression**: two SVM with linear and radial basis function (RBF) as the kernel.	Listener engagement classification reached 65% balanced accuracy
Mehta et al. ([141])	Use DAiSEE and Emoti-W dataset	Images	**Pre-processing**: Dlib face detector. **4-class classification and regression**: 3D CNN with a 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. ([68])	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. ([203])	Use existing dataset[1]	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**: pre-trained with pre-trained 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 [58].	**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
Shen et al. ([193])	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 facial expression recognition. A soft attention module (SE) was adopted to reduce the impact of the complex background.	Accuracy = 56%
Apicella et al. ([13])	EEG	EEg Signal	**Pipeline**: Filter bank, Common Spatial Pattern, SVM. **Pre-process**: artifact removal using independent component analysis (ICA), namely the 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.

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

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

PUBLICATIONS

(International Journal)

1. S. N. Karimah and Shinobu Hasegawa, "Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods," Smart Learning Environments, 9(1):31, 11 2022, doi: 10.1186/s40561-022-00212-y - *Refereed*

2. S. N. Karimah, H. Phan, Miftakhurrokhmat and S. Hasegawa, "Design Principle of an Automatic Engagement Estimation System in a Synchronous Distance Learning Practice," in IEEE Access, vol. 12, pp. 25598-25611, 2024, doi: 10.1109/ACCESS.2024.3366552 - *Refereed*

(International Conference)

1. S. N. Karimah and S. Hasegawa, "MeetmEE: Engagement Estimation-based Online Meeting Room for Distance Learning," in IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE) 2022, 4-7 December 2022, doi: 10.1109/TALE54877.2022.00090 - *Referred*

2. S. N. Karimah and S. Hasegawa, "A Real-time Engagement Assessment for Learner in Asynchronous Distance Learning," in The 17^{th} International Conference on Knowledge, Information and Creativity Support Systems (KICSS), November 23-25, 2022, doi: 10.52731/liir.v003.064 - *Referred*

3. S. N. Karimah, T. Unoki, and S. Hasegawa, "Implementation of Long Short-Term Memory (LSTM) Models for Engagement Estimation in Online Learning," in 2021 IEEE International Conference on Engineering, Technology & Education (TALE), Dec. 2021, pp. 283–289, doi: 10.1109/TALE52509.2021.9678909 - *Referred*

4. S. N. Karimah and S. Hasegawa, "Automatic Engagement Recognition for Distance Learning Systems: A Literature Study of Engagement Datasets and Methods," in International Conference on Human-Computer Interaction, 2021, pp. 264–276, doi: 10.1007/978-3-030-78114-9_19 - *Referred*

5. S. N. Karimah and S. Hasegawa, "A Real-time Engagement Assessment in Online Learning Process Using Convolutional Neural Network," in The 12th Asian Conference on Education (ACE2020), Jan. 2020, pp. 437–448, doi: 10.22492/issn.2186-5892.2021.39 - *Referred*

6. S. Hasegawa, A. Hirako, X. Zheng, S. N. Karimah, K. Ota, and T. Unoki, "Learner's Mental State Estimation with PC Built-in Camera," in Learning and Collaboration Technologies. Human and Technology Ecosystems, 2020, pp. 165–175, doi: 10.1007/978-3-030-50506-6_12 - *Referred*

(Grants)

1. Doctoral Research Fellowship (DRF), JAIST, October 2019 – March 2019.