AUTOMATICALLY DETECTING AFFECT IN COMPUTER-BASED LEARNING ENVIRONMENTS: A SYSTEMATIC LITERATURE REVIEW

by

Lydia Marion González Esparza

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation, and Data Science

Department of Educational Psychology
University of Alberta

Abstract

Affect detection is increasingly viewed as an essential component of computer-based learning systems because it aims to improve learner outcomes by adapting to the learner's affect. However, most computer-based learning environments used across formal and informal educational contexts do not respond to students' affective needs. Moreover, it is not clear which affective states should be assessed and which states have a positive or a negative effect on student learning. The aim of this review is to examine how affect is automatically detected and analyzed via affect-sensitive computational systems in educational settings. This systematic literature review analyzes 36 peer-reviewed publications that focus on finding relationships between affect and learning in computational applications. Evidence from the reviewed articles shows that most studies (1) were published in the last four years; (2) mainly used facial expressions to detect affect; (3) identified engagement, boredom, frustration, and confusion as the most frequent affective states in learning settings; and (4) used supervised machine learning algorithms to classify learner emotions. The present review identifies the following gaps in the related literature. First, it revealed that there is a paucity of studies in non-STEM domains and that sample K-12 students and participants from countries other than the US, given that two-thirds of the reviewed studies sampled university students, almost half of the studies sampled participants from North America, and almost three quarters of the studies focused on STEM contexts. Second, it identified facial expression as the most common physiological and behavioral data channel, with system log data being the most frequent performance-related channel. Third, it found that few studies examined both affect and achievement measures. Finally, it revealed that few studies employed unsupervised learning techniques or supervised learning regressors, given that supervised learning classifiers were overwhelmingly employed to predict affective states. This research provides recommendations on how to address these gaps, including the need for more methodological approaches, both theory- and data-driven, in capturing and analyzing affect. This review suggests the exploration and development of adaptive intelligent educational interfaces that use affective and behavioral states to provide a better learning experience by offering suitable responses. Likewise, the review suggests the exploration of creating affective datasets to improve existing machine learning affect detecting models.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Meaning
AC	Affective Computing
ACM	Association for Computing Machinery
AI	Artificial Intelligence
ASC	Autism Spectrum Condition
AUs	Action Units
ANN	Artificial Neural Network
BPMS	Body Pressure Measurement System
BROMP	Baker Rodrigo Ocumpaugh Monitoring Protocol
CBLE	Computer-based Learning Environment
CS	Computing Science
CV	Cross-validation
DBN	Dynamic Bayesian Network
EDA	Electrodermal Activity
EEG	Electroencephalogram
EKG/ECG	Electrocardiogram
ELA	English Language Arts
EMG	Electromyogram
EOG	Electrooculogram
ERIC	Education Resources Information Center
FACS	Facial Action Coding System
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy

GSR	Galvanic Skin Response
HART	Human Affect Recording Tool
HCI	Human-Computer Interface
IEEE	Institute of Electrical and Electronics Engineers
ITS	Intelligent Tutoring System
K-12	Kindergarten to 12th grade
KNN	K-Nearest Neighbor
LMMSE	Linear Minimum Mean Square Error
MAUI	Multimodal Affective User Interface
ML	Machine Learning
MOOC	Massive Open Online Courses
NN	Neural Network
PPEM	Psychophysiological Emotional Map
PPG	Photoplethysmography
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
RBF	Radial Basis Function Kernel
SMI	SensoMotoric Instruments
STEM	Science, Technology, Engineering, and Mathematics
SVM	Support Vector Machine

Chapter 1. Introduction

Modeling the user interaction experience is important for the design and uptake of adaptive intelligent systems. Affective user modeling has received increasing attention with the proliferation of digital devices, and it has been prompted by advances in the fields of human-computer interaction (HCI), cognitive science, psychology, education, neuroscience, and computing science. These advances have improved computational systems' detection of user behaviors, affect, and emotional reactions, which is important, given that emotions can have a considerable impact on learning experiences and subsequently on learning and performance outcomes.

In recent years, several models have been built to detect learning-centered affect (e.g., boredom, confusion, happiness, delight, motivation, engaged concentration, anxiety, and frustration) in computer-based learning environments (CBLE) with the ultimate goal of creating affect-sensitive, adaptive learning systems that improve learning outcomes. This has the potential to improve the learning experience by monitoring learners' progress and providing timely interventions. For example, if a student feels frustrated, the system may suggest an easier problem or revisiting a tutorial, whereas if a student seems engaged, the system may not intervene.

Modeling user experience is a challenging task mainly because the detection of affect is difficult to accomplish in computational systems (Jraidi et al., 2013). For example, frustration was shown to be difficult to both identify and address in intelligent tutoring systems (ITS) (DeFalco et al., 2018). Moreover, it is not clear which affective states have positive or negative effects on student learning (Bosch et al., 2015; D'Mello, 2013; Pekrun et al., 2007). For example, it was found that long-term frustration was negatively associated with learning outcomes, whereas short-term frustration was not problematic for learning (DeFalco et al., 2018; D'Mello & Graesser, 2009; Liu et al., 2017; Robison et al., 2009). Also, boredom has been negatively associated with learning outcomes (Craig et al., 2004), whereas engaged concentration has been positively associated with learning outcomes (Pardos et al., 2014).

The use of affective computing within a system is intended to meet one or more of the three possible goals that were described by Picard: 1) to detect user emotions; 2) to express a human emotion (e.g., an avatar, robot, and animated conversational agent); and 3) to "feel" an emotion (Calvo & D'Mello, 2010; Picard, 1997). An interdisciplinary literature review in the field of affective computing (AC) examined affect-detection systems from the perspective of several key emotion theories, methods, and data sources (Calvo & D'Mello, 2010).

1.1 Challenges

Affect detection is difficult because emotions cannot be measured directly, and they vary in both expression and experience from one individual to another. Overlapping areas between emotion research (i.e., affective science) and affective computing include affect expression and detection by humans and computers (Calvo & D'Mello, 2010). Learning-centered affective states are different from the basic emotions whose relationships with expressions were thoroughly studied for decades. Thus, it is not clear whether there are similar links between learning-centered affective states and expressions (Bosch et al., 2015). Moreover, using physical sensors to collect data about learners' affective states has one important limitation in that it is preferred if studies are conducted in a controlled laboratory setting, which hinders the generalizability of the findings stemming from this research (Baker & Ocumpaugh, 2014). Even using the more unobtrusive interaction-based affect detection may limit the generalization of the interaction-based detectors across populations and systems (Kai et al., 2015). The present review aims to identify any potential gaps in knowledge about affect detection.

1.2 Study Purpose

This review examines the relevant literature spanning several and often overlapping fields concerned with the theory, methodology, and practice of detecting affect. One of the goals of the present literature review is affect detection (i.e., the detection of individuals' affective states, including emotions, feelings, moods, attitudes, affective styles, and temperament; Calvo & D'Mello, 2010) in the broader context of emotion research (i.e., affective science) and affective computing. The current review aims to ascertain the degree to which computational systems can automatically recognize or respond to users' affective states. Moreover, the review aims to investigate whether affect-sensitive interfaces facilitate human-computer interaction in terms of enjoyment and effectiveness (e.g., learning gains). The review poses the following research questions:

RQ1: What are the key characteristics of the studies reviewed?

RQ2: What are the channels of data employed in the studies reviewed?

RQ3: What are the affective states (classes of emotions) investigated in the studies reviewed?

RQ4: What are the machine learning algorithms used to detect affect in the studies reviewed?

Chapter 2. Conceptual Background

This section starts with the definition and theories of emotions found in the literature, followed by the definition of affect and its relationship with learning. Next, it describes the kinds of data modalities and how these are used for affect detection. Afterwards, it introduces one of the most commonly found affect classification observation method. Moreover, it discusses how machine learning techniques can be used to predict and classify learners' affective states. Further, it discusses the Cultural Dimensions Theory and delves into how affect can vary across cultures. Finally, it describes the physical data channels explored in this review.

2.1 Theories of Emotion

Emotions are usually conceptualized using six perspectives: expressions, embodiments, outcomes of cognitive appraisals, social constructs, products of neural circuitry activity, and psychological interpretation of basic feelings or core affect (Calvo & D'Mello, 2010). The first four perspectives (i.e., expressions, embodiments, outcomes of cognitive appraisals, and social constructs) draw on traditional emotion theories (Calvo & D'Mello, 2010). The fifth perspective (i.e., products of neural circuitry activity) draws on theories related to affective neuroscience. Finally, Russell (1980) introduced a unified theory of emotion to bridge the gaps from the previous theories by proposing that emotions are neurophysical states with either a positive or negative valence and a level of arousal.

2.2 Affect

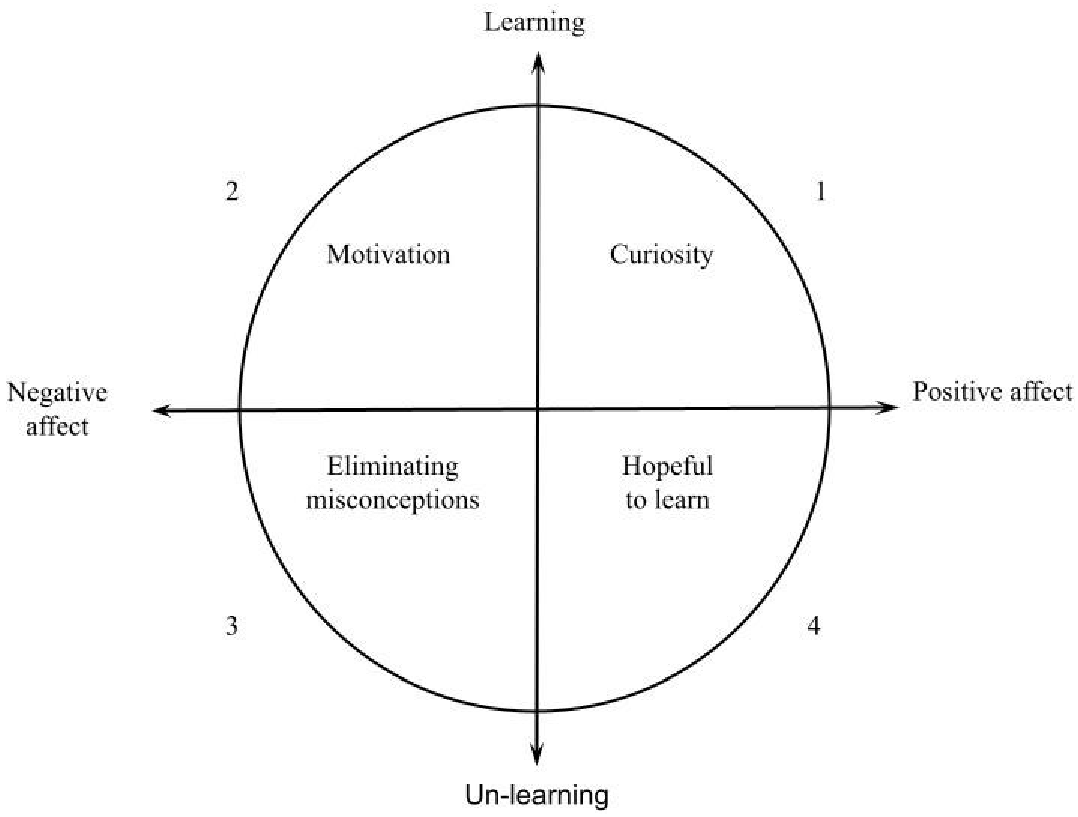
Psychologists describe affect as a set of independent dimensions that can range from positive to negative. Nowlis and Nowlis (1956) concluded that there are at least six factors of affect: sadness, anxiety, anger, elation, tension, and preference. Russell (1980) elaborates on Schlosberg's (1952) proposal of these six factors of affect, posing that, rather than being independent of each other, these factors are related to each other, and when illustrated, they should be organized in a circle representing two bipolar dimensions, rather than at least six monopolar ones.

2.3 Learning and Affect

Kort et al. (2001) view learning as a cyclical process that involves many affective responses as it often entails various levels of natural failure. This failure can lead to a lack of motivation and can occasionally lead learners to abandon an activity. Therefore, accurately identifying the learner's affective state during the learning process is important to avoid attrition and keep motivation levels high, especially in science, math, engineering, and technology (STEM) learners. The authors propose a model where the relationship between affective states and the learning process is associated. In a circular manner and divided into four quadrants that are numbered in a counterclockwise fashion, being the first (i.e., curious about the subject) and second (i.e., motivated to reduce confusion) quadrant where ideally the learner starts the process. After a failure, it is expected that the learner will fall to the third quadrant (i.e., frustrated and eliminating some misconceptions). When the learner has reviewed the mistakes, they may move to the fourth quadrant (i.e., hopeful about learning the subject). As it is a cyclical process, a new idea or adequate feedback can help the learner jump to the first quadrant again, as shown in Figure 1. According to this theory of affect, it is important to approach the learner when they are in the quadrants related to negative affective states and negative learning experiences and help them see that these emotions are part of the process.

Figure 1

Graphical Representation of the Cyclical Process of Learning Involving Affective Responses Based on Kort, Riley, and Piccard (2001)



2.4 Sensors

A sensor is a device that detects and records a physical electronically transmitted input and produces an output of the measurements of the detected phenomenon (Hao & Foster, 2008). Sensors can function as a measuring device (e.g., electrodermal activity sensor) for a phenomenon or respond to the measurement of the signal (e.g., motion-activated light).

2.5 Data Modalities: Data Channels for Affect Detection

2.5.1 Data Modality

A data modality (i.e., signal) is a single input or output of data obtained by a channel between a human and a computer (Karray et al., 2007). Data modalities could be used by themselves (unimodally) or in combination (multimodally) to obtain richer information about the subject that is being sensed.

Data modalities of affect-detection techniques can be classified into more long-established modalities (e.g., physiology, facial expressions, voice), burgeoning modalities (e.g., brain imaging/neuroimaging, text processing, body language, posture), and combinations of these modalities (Calvo & D'Mello, 2010).

For example, in intelligent tutoring systems, learner affect has been detected using less invasive interaction, visual, and audio data, including webcams to capture facial expressions, gaze trackers to capture eye-gaze patterns, facial expressions (Lan et al., 2020), but also physiological-dependent wearable devices that measure physiological signs through physiological sensors such as pressure-sensitive devices (e.g., seats, back pads, keyboards and mice, and posture-sensing chairs), skin conductance sensors or wristbands (i.e., electrodermal activity or EDA; or galvanic skin response or GSR), electroencephalogram (EEG), functional near-infrared spectroscopy (fNIRS), and heart rate (PPG; Alqahtani et al., 2021). These types of data are usually collected to infer learners' cognitive states (e.g., attention, memory workload; Dorneich et al., 2007), affective states (e.g., boredom, confusion, engagement, frustration; Saxena et al., 2020), and behavioral states (e.g., on task, off task; Ding et al., 2022) during the learning process. Physical sensor-based detectors directly capture embodied representations of learners' affective states via multiple data channels (e.g., facial expressions and body movements or posture via a webcam). For instance, vision-based affect detection can be inexpensive, non-invasive, and is available on many devices (e.g., a webcam is a common vision-based sensor), facilitating affect detection through facial recognition or body movements. In contrast, interaction-based detectors indirectly capture learners' affective states via learners' actions within the computer-based system (e.g., the time between the start and the end of the interaction, the number of actions taken during the interaction with the system, speed of help requests). For this reason, physical sensor-based detectors often outperform interaction-based detectors in predicting some affective states, including delight, engaged concentration, and frustration (Kai et al., 2015). However, the physiological sensor-based detectors are applicable in more learning contexts than the former detectors. Affect detectors have been developed to collect data via multiple data types, including physiological, behavioral, and performance related.

2.5.2 Data Channel

A data channel refers to a medium that transports and delivers data (Shim et al., 2022). A data channel occurs where a reaction caused by a human produces data bits that are encoded by a computer-based sensor and later decoded by a software to retrieve human-readable information about that interaction. For example, the Empatica E4 wristband is composed of several sensors, each of them collecting data from different channels, such as temperature, voltage level of electrodermal activity, and blood-volume pressure.

2.5.2.1 Physiological Channels

Physiological features track bodily changes associated with emotion. For example, galvanic skin response (GSR) or electrodermal activity is usually linked with emotional arousal (Chatterjee et al., 2022). Cardiovascular measures, such as heart rate (HR), are employed to understand the autonomic nervous system function and are associated with emotional valence (Griffin & Howard, 2022). Brain activity collected through electroencephalograms (EEG) provides neural indexes related to cognitive changes such as alertness, attention, workload, executive function, or verbal and spatial memory (Buzsáki & Watson, 2012). System log data is employed to understand the engagement the learner is experiencing in the platform, such as clickstream patterns. A study found that clickstream patterns combined with task performance were associated with frustration and boredom (Yu et al., 2019). Yue et al. (2019) used a combination of webcam video, eye tracking data, and clickstream data to detect engagement in students learning Python programming in a MOOC.

2.5.2.2 Behavioral Channels

Behavioral features track the interaction between the learner and the learning environment to ascertain the degree of involvement with the task (e.g., rate of requesting help, hints used, mouse or keyboard presses, click frequency). System log data is also used to capture behavior as it provides a registry of the interaction between the learner and the environment. This rich data channel is usually composed of the event information (e.g., timestamp, mouse location) and other behavioral markers (e.g., number of hints requested, timestamp of requested hint, or whether a task was finished or not).

2.5.2.3 Performance-Related Channels

Performance features constitute objective measures of the level of task proficiency (e.g., correctness, errors made in the task, response time). These features are useful when predicting academic performance and when combined with sensor-based data that may be able to create generalizable models for affect detection. For example, the system logs record the number of attempts before completing a task or whether a task was completed correctly or incorrectly. Rajendran et al. (2019) used system log data to compare how motivational messages influence the degree of frustration detected in students when learning mathematics. There are other variables that are useful when interacting with the computational system, such as the context (i.e., environment, such as task difficulty, the relevance of the hints or help provided, the imposed time constraints) and the learner profile (i.e., user characteristics, such as the learner's goal, preference, skills, personality, computer usage frequency). *Physiological sensor-based detection* infers affective states from the physical reactions of the learner (e.g., video-based data collected via a webcam). *Interaction-based detection* infers affective states from interactions of the learner with the computer-based learning system (e.g., logs of the learners' interaction with the learning environment). *Performance-related detection* infers affective states from task performance of the learner (e.g., scores, error rate, number of hints requested, etc.).

2.6 Observation Methods

The *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP 2.0)* is a field coding protocol. Researchers use BROMP to perform live identification of predominant affective states (e.g., boredom, confusion, engaged concentration, frustration, delight, and dejection) and behavioral states (on task, on-task conversation, and off-task) of learners while they perform a learning activity (Bosch et al., 2015). It has been used to observe each learner individually until visible affect is detected or 20 seconds have passed; then, the researchers focus on another student, in a round-robin manner (Baker & Ocumpaugh, 2014). Usually, BROMP is implemented using the Android app, *Human Affect Recording Tool* (HART; Baker et al., 2012). Furthermore, BROMP is then used to support supervised learning to build affect and behavioral state detectors.

2.7 Self-report Measures

Self-report is a test-taking method in which participants provide a rating on a certain scale of their own characteristics (e.g., mindset or confidence; Johnson & Christensen, 2014). Self-report observation methods were found in the literature as an alternative, an addition, or a confirmation of other observation methods, such as inventories. For instance, Grafsgaard et al. (2014) explored how models using different channels of data performed when predicting normalized learning gains. The model using a combination of the self-report and physiological data outperformed the models that only used unimodal data.

For example, the reviewed publications have employed a wide variety of self-report instruments. *Big Five Inventory* (BFI) is a self-report inventory designed to assess learners' personality traits, namely openness, conscientiousness, extraversion, agreeableness, and neuroticism (John et al., 1991). The *Intrinsic Motivation Inventory* (IMI) is a scale for measuring the sense of presence experienced in a virtual environment (Ryan, 1982). The *Presence Questionnaire* is a scale for measuring the sense of presence experienced in a virtual environment (Witmer & Singer, 1998). The *New General Self-Efficacy Scale* is an eight-item measure that assesses how much people believe they can achieve their goals, despite difficulties (Chen et al., 2001). The *Game Engagement Questionnaire* (GEQ) is used to quantify engagement of participants in games and includes a set of 19 questions classified into four categories: absorption, flow, presence, and immersion (Brockmyer et al., 2009). The *Self-Assessment Manikin* (SAM) survey is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli (Bradley & Lang, 1994). The *NASA-TLX Load Index* is a tool for measuring and conducting a subjective mental workload (MWL) in an assessment (Hart & Staveland, 1988). Finally, the *User Engagement Survey* (UES) measures self-reported user engagement (O'Brien et al., 2018). Some publications also used surveys to assess the quality of the activity and to receive feedback on the computer-based system.

2.8 Machine Learning

One of the most important components of an affect-sensitive system is the capacity to predict the affective state of students. Researchers have attempted to identify learners' affective states in computer-based learning environments by building machine learning models to categorize learners' emotional states. The integration of multimodal affect-sensing systems with artificial intelligence (AI) techniques and, particularly, with machine learning (ML) algorithms, may improve these systems' ability to infer learners' affective states. Classification models have been created to use features extracted from different modalities of data collection to predict several affective and behavioral states. For instance, classification algorithms used in affect detection (i.e., to predict emotional states) based on manifesting features of affect include rule-based reasoning, support vector machines, and tree-based models. Then, machine learning classifiers can be trained to recognize learner affect in learners' facial expressions or posture (e.g., confusion) and the computer-based system may intervene in real time by providing hints or by suggesting tutorials.

2.9 Cultural Dimensions Theory

Culture is the societal programming that helps individuals distinguish from different members of society (Hofstede, 2011). Cross-cultural communication is essential for society and economic benefits in contemporary society. Hofstede developed the cultural dimensions theory framework to show how a society's culture influences the value of its members and their behavior. Although it was first proposed around 1970, this theory has been reworked and updated due to new findings in research. The current Hofstede model proposes six dimensions of culture that can be analyzed: (1) power distance, measures how a culture views relationships of power between people, (2) uncertainty avoidance, measures the extent to which a culture attempts to cope with anxiety of an unknown future, (3) individualism versus collectivism, measures the degree to which a culture is integrated into groups, (4) masculinity versus femininity, measures the extent to which a culture's gender roles are distinguished, (5) long term versus short term orientation, measures a cultures choice of focus on the past, present and future, and (6) indulgence versus restraint, measures the degree to which a culture allows basic human desires to conduct behavior.

Literature indicates that the way and the extent to which we display emotions are influenced by our culture (Kleinsmith et al., 2006). Although body posture and non-verbal behavior has been found to vary cross-culturally (Davidson et al., 2009), facial expressions for some emotions have been found to be similar across cultures (Mandal & Ambady, 2004).

2.10 Affect Detection: Signals, Data, and Methods

Calvo and D’Mello’s (2010) review discussed affective computing approaches to affect detection, linking each approach to the closest corresponding emotion theory.

2.10.1 Facial Expressions

Most of the affect detection research detects basic emotions from the face, drawing on the “emotions as expressions” view of modeling affect, which links basic emotions with distinctive facial expressions. The Facial Action Coding System (FACS; Ekman & Friesen, 1978) offers a more standardized method to classify facial activity using “facial actions” (i.e., facial motions) to identify six basic emotions: anger, disgust, fear, joy, sadness, and surprise (Ekman, 1992). This system enables human coders to decompose an expression into action units (AUs) that identify independent facial motions. The coding of emotions is a tedious and time-consuming task, which prompted several alternative techniques based on machine learning classifiers to automatically detect AUs from static images (Calvo & D’Mello, 2010).

2.10.2 Body Language and Posture

This affect-detection modality provides valuable and reliable information not captured via other modalities that are prone to social editing. For instance, automated temporal-transition posture analysis in a learning environment was conducted using the Tekscan’s Body Pressure Measurement System (BPMS) to infer affective states (e.g., interest in a task) of children during a computer learning task (Mota & Picard, 2003). Similar to the analyses used for other modalities (e.g., face and voice), a machine learning classifier (i.e., neural network) was used to analyze posture sequences over a 3-second interval and predicted nine static postures with high accuracy (Mota & Picard, 2003). Machine learning techniques were also used to detect boredom, confusion, delight, flow, and frustration, respectively, with accuracy ranging from 70 to 83 percent (D’Mello & Graesser, 2009).

2.10.3 Physiology

AC applications analyze physiological activity to extract affect patterns that capture the expression of emotions, drawing on theories of embodiment of emotion. The measures employed to capture physiological states are based on electrical signals produced by the brain (e.g., Electroencephalography or EEG), heart (electrocardiogram or EKG/ECG), muscles (e.g., electromyogram or EMG), eye movement (e.g., electrooculogram or EOG), and skin (e.g., electrodermal activity or EDA). Researchers devised the multimodal affective user interface (MAUI) system to map physiological signals to emotions (Nasoz et al., 2004). A Psychophysiological Emotional Map (PPEM) was proposed to link physiological signals (e.g., heart rate and skin conductance) to dimensional models (e.g., valence, arousal) through user-dependent mappings (Villon & Lisetti, 2006).

2.10.4 Brain Imaging and EEG

Affective neuroscience has focused on mapping neural circuitry corresponding to emotion using techniques such as Functional Magnetic Resonance Imaging (fMRI) or Electroencephalography (EEG).

2.11 Computer-based Learning Environments

Computer-based learning environment (CBLE) refers to the wide variety of technologies employed to support learning grounded in learning theories (Lajoie & Naismith, 2012).

For example, an Intelligent Tutoring System (ITS) is a CBLE that must have: (1) domain knowledge (expert model); (2) knowledge of the learner (student model); (3) knowledge of teaching strategies (tutor) (Shute & Psotka, 1994). ITSs were developed to enhance what or how much a student is learning through engaging mechanisms to personalize teaching (Joshi et al., 2019). Examples of ITSs include MetaTutor, AutoTutor, and MathSpring. Students use the ITS to complete activities to stimulate learning, whereas the instructor uses it to understand and adapt teaching strategies (Han et al., 2019).

Massive Open Online Courses (MOOC) are web-based learning programs designed for large groups of students (Terras & Ramsay, 2015). Students can register for free or at a low cost to these programs and learn at their own pace (Zhu et al., 2020). Examples of MOOCs include Harvard University's Introduction to Computer Science course hosted on the edX platform and Stanford University's Machine Learning course hosted by Coursera.

An online platform is a tool where online services such as websites and mobile applications are hosted (Sofi Dinesh et al., 2021). Examples of online platforms include Amazon Web Services, Microsoft Azure, and Google Cloud Platform. In education, an online platform refers to a portal where educational resources are stored in one place (Nurhudatiana et al., 2018).

A Learning Management System (LMS) is a platform, usually hosted online, for the tracking, reporting, and delivery of educational courses usually found in university settings (Oliveira et al., 2016). Some examples of LMSs are Moodle, Blackboard Learn, and Schoology.

Chapter 3. Method

3.1 Search Strategy

A comprehensive strategy to search for relevant records guided the present review to ensure the identification of a wide range of studies. Publications for this systematic literature review were gathered from ACM Digital Library, ERIC, IEEE Xplore, and SpringerLink. These databases were selected as they contain an extensive array of publications on education and technology and are representative of the work published in the topic of interest, in this case the relationship of affective state with learning, for the present review. The Association for Computing Machinery (ACM) Digital Library is an online archive of computer science work from the 1950s onwards. It contains the full text to ACM journals, magazines, conference proceedings, and e-books on topics related to computers and technology. ERIC is a comprehensive educational database that includes studies from 1996 onwards that cover pre-school to post-secondary and adult education. IEEE Xplore covers electrical and electronic engineering, computer engineering, and computing science topics; it provides full-text access to IEEE transactions, journals, magazines, and conference proceedings published since 1884. Springer Link is a database that provides access to the full text of books and journals from Springer-Verlag and associated publishers, covering life sciences, chemical sciences, geosciences, computer science, mathematics, medicine, physics, astronomy, engineering, environmental sciences, law, and economics.

The searches were conducted using the following Boolean search terms: “*affect*” AND (“*sensor*” OR “*physiologic*” OR webcam) AND “learn*” AND (“programming” OR "gam*" OR “intelligent tutoring system*”) AND “student*”. After the retrieval of peer-reviewed publications (i.e., journal articles and conference proceedings), 542 studies (478 retrieved from the ACM Digital Library, 26 from ERIC, 38 from IEEE Xplore, and 2 from SpringerLink) were imported into the Covidence (2019) systematic review platform for screening. Of the total studies, 17 duplicates were removed during screening. One rater completed the process, and a second rater reviewed the steps of the process leading to the selection of the final set of records. Then, two other raters reviewed independently the summaries of the final records included. Only records published until June 2021 were selected.

3.2 Exclusion Criteria

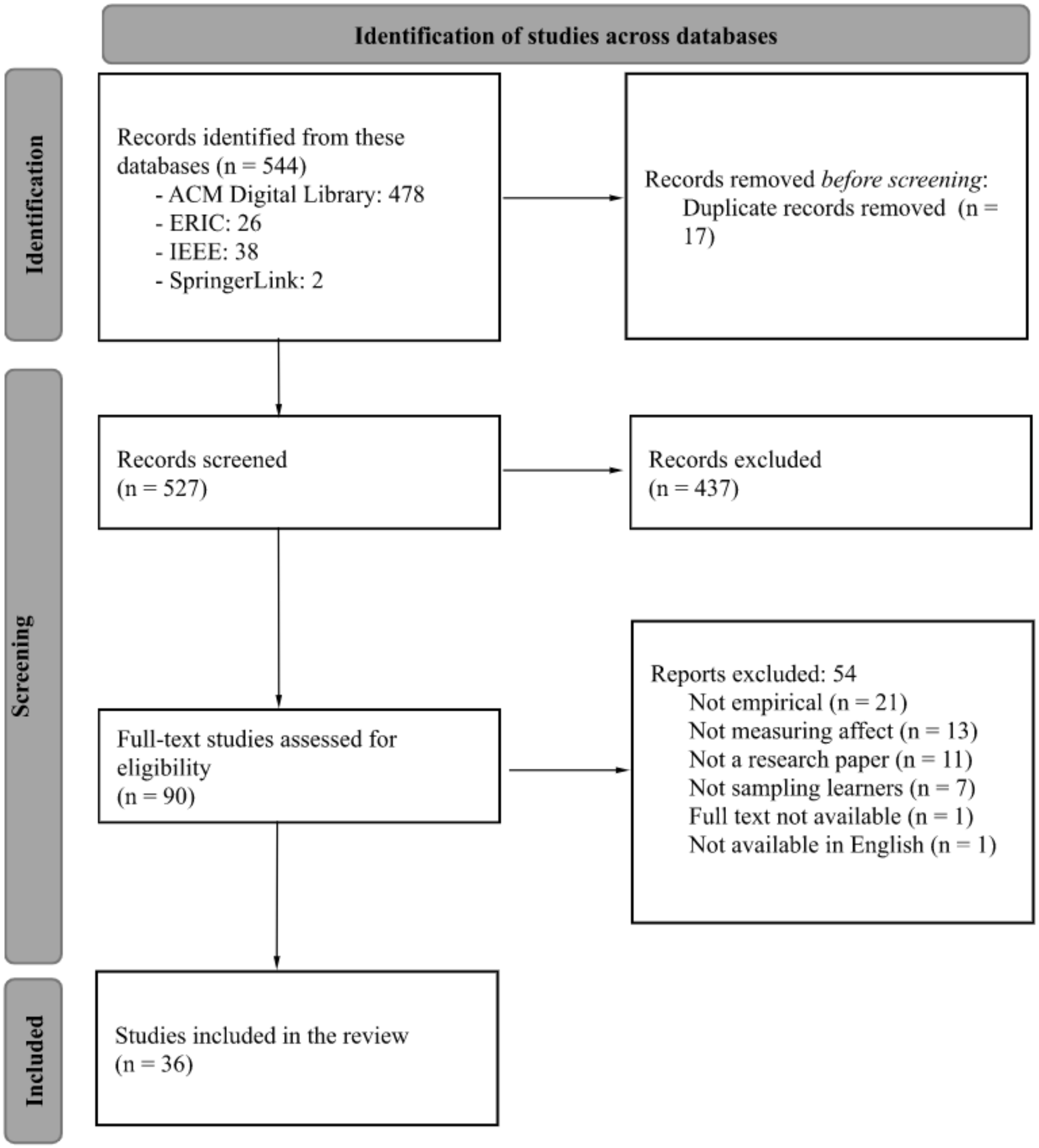
This systematic literature review follows the PRISMA (Page et al., 2021) guidelines for article selection. This process was facilitated by Covidence (2019). The first stage was the screening stage of the title and abstract, where one of the raters removed the publications that were not eligible for inclusion. This process required the reviewer to critically analyze the title and abstract of each publication to determine if the topic and procedures aligned with the purpose of this literature review. During this stage, 435 records were discarded. Publications that used the word "affect" referring to causality were discarded as this was not the definition of interest. The keyword "sensors" yielded results in the computer systems area, which were also discarded as the type of sensors used did not capture any physiological data. Project proposals were also discarded as they did not provide evidence of the claims. Finally, publications that only described the procedure of the intervention and not the results were also discarded.

The following stage was the full-text review. We excluded publications that (1) were not empirical studies, (2) were not peer-reviewed articles or conference proceedings, (3) were not available in English, (4) lacked the full text, (5) were not sampling learners, and (6) were not measuring affective state. Only peer-reviewed publications were retained, as they tend to include more rigorous methodologies, including more methodological detail and more robust results.

During this stage, 54 records were discarded. The final number of publications included for this review was 36 records, as shown in Figure 2.

Figure 2

The PRISMA 2020 Flow Diagram for New Systematic Reviews Employed in This Study


Identification of studies across databases
Identification
Screening
Included
Records identified from these databases (n = 544)
- ACM Digital Library: 478
- ERIC: 26
- IEEE: 38
- SpringerLink: 2
Records removed *before screening*:
Duplicate records removed (n = 17)
Records screened (n = 527)
Records excluded (n = 437)
Full-text studies assessed for eligibility (n = 90)
Reports excluded: 54
Not empirical (n = 21)
Not measuring affect (n = 13)
Not a research paper (n = 11)
Not sampling learners (n = 7)
Full text not available (n = 1)
Not available in English (n = 1)
Studies included in the review (n = 36)

3.3 Coding Procedure

During the extraction phase, all 36 records were thoroughly screened for ensure they provide data to answer the research questions. Table 1 describes the variables used in this review. Variables were divided into two categories: characteristics (i.e., those variables that identify the publications) and methodology (i.e., those variables that were employed in the experiments of the reviewed publications).

Table 1

Description of the Variable Extracted from the Reviewed Publications

	Variable	**Description**
Characteristics	Title	Title of the record
	Authors	Full names of the authors of the record
	Venue	Venue where the record was published
	Year	Year when the record was published
	Author Country	Country of the university the author is associated in the record
	Participant Country	Reported country of the participants of the experiment
	Educational Level	Level of education reported of the participants of the experiment
	Research Design	Research design employed in the experiment
Methodology	Domain	Educational domain of the experiment
	System Type	Type of computer-based system employed in the experiment
	Sensor	Sensor or sensors employed to record data channels in the experiment
	Observation method	Methods employed in the experiment to infer or define the participants' affective state in the experiment
	Emotion	Affective state or states recognized in the experiment

	ML Algorithm	Machine learning algorithms employed to analyze data from the experiment

3.4 Inter-rater Reliability

One rater reviewed all 36 publications and another rater reviewer 15% of these publications. To find the inter-rater reliability coefficient, the records that were coded by both authors were compared. After both raters extracted the necessary data from the records, the results were coded into categories. The data was transformed into two vectors and analyzed using IBM SPSS Statistics. The Cohen's Kappa coefficient, κ, that represents the inter-rater reliability was $\kappa = 0.929$ (95% CI, 0.876 to 0.973), $p < 0.001$, indicating near perfect agreement.

Chapter 4. Results

A summary of the key characteristics of each reviewed publication is included in Table 2, showing the publication identifier, citation in APA format, publication year, geographical location of the authors and of the study sample, educational level, and research design. Table 3 includes the methodology employed in each of the reviewed publications, subject domain of the study, the application (i.e., educational game, intelligent tutoring system) used in the experiment, the sensors used to collect affective data, and the machine learning algorithm applied to analyze the data. Table A1 in the *Appendix A* shows all the publications included in the present systematic review.

Table 2

Key Characteristics of the Reviewed Publications

Pub ID	APA Citation	Year	Database	Author Country	Participant Country	Education Level	Research Design
1	Barron-Estrada et al. (2018)	2018	IEEE	Mexico	Mexico	University	Experimental
2	Bosch et al. (2015)	2015	ACM	USA	USA	K-12	Quasi-Experimental
3	Burleson & Picard (2007)	2007	IEEE	USA	USA	K-12	Experimental
4	DeFalco et al. (2018)	2018	Springer Link	USA	USA	University	Experimental
5	Ghaleb et al. (2018)	2018	IEEE	Netherlands	Netherlands	University	Correlational
6	Grafsgaard et al. (2014)	2014	ACM	USA	USA	University	Quasi-Experimental
7	Joshi et al. (2019)	2019	IEEE	USA	Not disclosed	University	Correlational
8	Jraidi & Frasson (2013)	2013	ERIC	Canada	Not disclosed	University	Experimental
9	Jraidi et al. (2013)	2013	ACM	Canada	Not disclosed	Not disclosed	Experimental

10	Kai et al. (2015)	2015	ERIC	USA	USA	K-12	Quasi-Experimental
11	Korn & Rees (2019)	2019	ACM	Germany	Germany	University	Experimental
12	Lee-Cultura et al. (2020)	2020	ACM	Norway, Greece	Norway	K-12	Quasi-Experimental
13	Leong (2015)	2015	IEEE	UK	Singapore	University	Experimental
14	Mangaroska et al. (2020)	2020	ERIC	Norway, Australia	Norway	University	Experimental
15	Mills et al. (2017)	2017	ACM	Canada	USA	K-12	Quasi-Experimental
16	Muñoz et al. (2016)	2016	ERIC	Mexico, Ireland, UK	Mexico	University	Quasi-Experimental
17	Park et al. (2018)	2018	ACM	South Korea, USA	Not disclosed	University	Quasi-Experimental
18	Pham & Wang (2016)	2016	ACM	USA	USA	University	Experimental
19	Psaltis et al. (2017)	2018	IEEE	Greece	Greece	K-12	Experimental
20	Rajendran et al. (2019)	2019	IEEE	India	India	K-12	Quasi-Experimental
21	Sharma et al. (2018)	2018	ACM	Norway	Norway	K-12	Experimental
22	Sharma et al. (2021)	2021	ACM	Norway	Norway	K-12	Experimental
23	Sinha et al. (2015)	2015	IEEE	India	India	Professionals	Experimental
24	Sottilare & Proctor (2012	2012	ERIC	USA	USA	University	Quasi-Experimental
25	Srivastava et al. (2018)	2018	ACM	Australia	Australia	Professionals	Quasi-Experimental
26	Standen et al. (2020)	2020	ERIC	UK, Italy, Spain	UK, Italy, Spain	K-12	Quasi-Experimental
27	Subburaj et al. (2020)	2020	ACM	USA	USA	University	Experimental
28	The & Mavrikis (2016)	2016	ACM	Singapore, UK	Singapore	University	Experimental
29	Vail et al. (2016)	2016	ACM	USA	USA	University	Quasi-Experimental
30	VanLehn et al. (2017)	2017	IEEE	USA	USA	K-12 and University	Experimental
31	Veliyath et al. (2019)	2019	ACM	USA	USA	University	Experimental
32	Wu et al. (2020)	2020	Springer Link	Taiwan	Taiwan	University	Experimental

33	Xiao & Wang (2015)	2015	ACM	USA	USA	University	Experimental
34	Xiao & Wang (2016)	2016	ACM	USA	USA	University	Experimental
35	Yang et al. (2019)	2019	ERIC	USA	Not disclosed	K-12	Experimental
36	Yue et al. (2019)	2019	IEEE	China, UK, USA	China	University	Experimental

Note: PubID = Publication Identifier

Table 3

Methodology Employed in the Reviewed Publications

Pub ID	**APA Citation**	**Domain**	**System Type**	**Sensor**	**ML Algorithm**
1	Barron-Estrada et al. (2018)	English to Spanish Language	Online platform	Cellphone: microphone, accelerometer, and gyroscope.	SVM
2	Bosch et al. (2015)	Physics	Game	Webcam	SVM, C4.5 trees, and Bayesian classifiers.
3	Burleson & Picard (2007)	Mathematics	Game	Video camera, a pressure mouse, a skin-conductance sensor, and a posture chair	Not disclosed
4	DeFalco et al. (2018)	Military Medical Training	ITS	Microsoft Kinect and Affectiva Q-Sensor	J48, JRip, Logistic, Regression, Naïve Bayes, SVM, Step Regression, KStar
5	Ghaleb et al. (2018)	Mathematics, History, Sports, and Geography	Game	Mouse and keyboard	SVM
6	Grafsgaard et al. (2014)	Programming	ITS	Kinect camera, webcam, skin conductance sensor, and database logs	J48 DT
7	Joshi et al. (2019)	Mathematics	ITS	Webcam and GoPro, mouse location and clickstream, video stream of the screen.	Multi-layer perceptron with Adam optimizer
8	Jraidi & Frasson (2013)	Mathematics	ITS	Electroencephalogram, skin conductance sensor, and blood volume pulse sensor	SVM, Decision tree, Naïve Bayes
9	Jraidi et al. (2013)	Mathematics	ITS	EEG, video cameras and mouse movement	DBN

10	Kai et al. (2015)	Physics	Game	Webcam	JRip, J48 decision trees, KStar, Naïve Bayes, step and logistic regression
11	Korn & Rees (2019)	Mathematics	Game	Video cameras and Shimmer sensor	Not disclosed
12	Lee-Cultura et al. (2020)	English Language	Game	Eye tracker. webcam, Empatica E4, and Kinect Skeleton	Not disclosed
13	Leong (2015)	Programming	NA	Webcam, keyboard stream, and screen video stream	Logistic regression with lasso regularization
14	Mangaroska et al. (2020)	Programming	Online platform	Empatica E4, eye tracker, and webcam	Random Forest
15	Mills et al. (2017)	Biology	ITS	QUASAR	Qstate classifier
16	Muñoz et al. (2016)	Physics	Game	Not disclosed	Bayesian Networks
17	Park et al. (2018)	Programming	Online platform	Platform logs	Not disclosed
18	Pham & Wang (2016)	Law	ITS	Attentive Review and webcam	Linear kernel ranking SVM
19	Psaltis et al. (2017)	Game Theory	Game	Kinect camera	ANN
20	Rajendran et al. (2019)	Mathematics	ITS	Not disclosed	Not disclosed
21	Sharma et al. (2018)	Programming	Online platform	SMI and eye tracker	Not disclosed
22	Sharma et al. (2021)	Programming	Online platform	Webcam	Not disclosed
23	Sinha et al. (2015)	Not disclosed	Game	Neurosky, pulse oximeter, eSense, game videostream, and mouse keystrokes	Gaussian Mixture Model
24	Sottilare & Proctor (2012	Military Medical Training	Game	Not disclosed	Not disclosed
25	Srivastava et al. (2018)	Programming	NA	Eye tracker	SVM, KNN, Random Forest
26	Standen et al. (2020)	Social skills, STEM learning, language, mathematics	Online platform	Audio, posture-sensing chair, and video camera	Linear mixed model and log-linear mixed models
27	Subburaj et al. (2020)	Physics	Game	Webcam and eye tracker	Random forest
28	The & Mavrikis (2016)	Programming	Online platform	Eye tracking device, video screen capture	Not disclosed
29	Vail et al. (2016)	Programming	ITS	Kinect camera, webcam, and skin conductance bracelet	Not disclosed

30	VanLehn et al. (2017)	Mathematics	ITS	Video camera and posture-sensing chair.	Not disclosed
31	Veliyath et al. (2019)	Mechanical engineering	NA	Eye tracker	Random Forest, SVM, Adaptive Boosting, and Extreme Gradient Boosting
32	Wu et al. (2020)	Physics	Game	Eye tracker and emWave	Not disclosed
33	Xiao & Wang (2015)	Game Theory	MOOC	Nexus 5 flashlight	RBF SVM
34	Xiao & Wang (2016)	STEM learning	MOOC	Neurosky Mindwave EEG headset and Nexus 5 flashlight	Ranking SVM
35	Yang et al. (2019)	Mathematics	ITS	Not disclosed	LMMSE
36	Yue et al. (2019)	Programming	MOOC	Microsoft LifeCam, eye-tracker, and clickstream	VGG16, Inception-ResNetV2, VGG16 with LSTM, Inception-ResNetV2 with LSTM, CART, Random Forest, GBDT

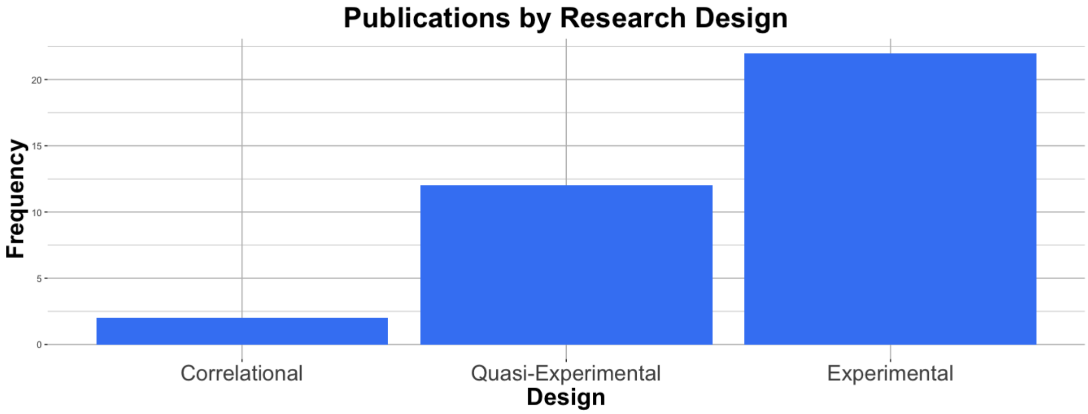
Note: PubID = Publication Identifier

4.1 What are the key characteristics of the studies reviewed?

Most research designs in the reviewed publications were experimental (n = 22; 61.11%) followed by quasi-experimental (n = 12; 33.33%) and correlational (n = 2; 5.55%), as shown in Figure 3. There were no studies employing a qualitative research design.

Figure 3

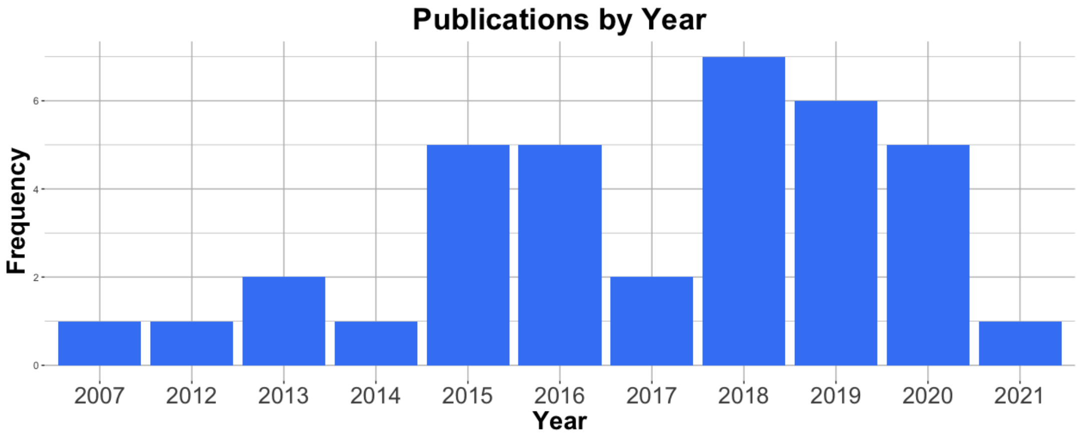
The Reviewed Publications by Research Design



More than half of the records reviewed were published in the last 4 years, with a peak in 2018 (the highest frequency year of publication with 7 publications), as shown in Figure 4. The results also show an overall increasing trend (with the exception of 2017), with 86% (n = 31) records being published since 2015.

Figure 4

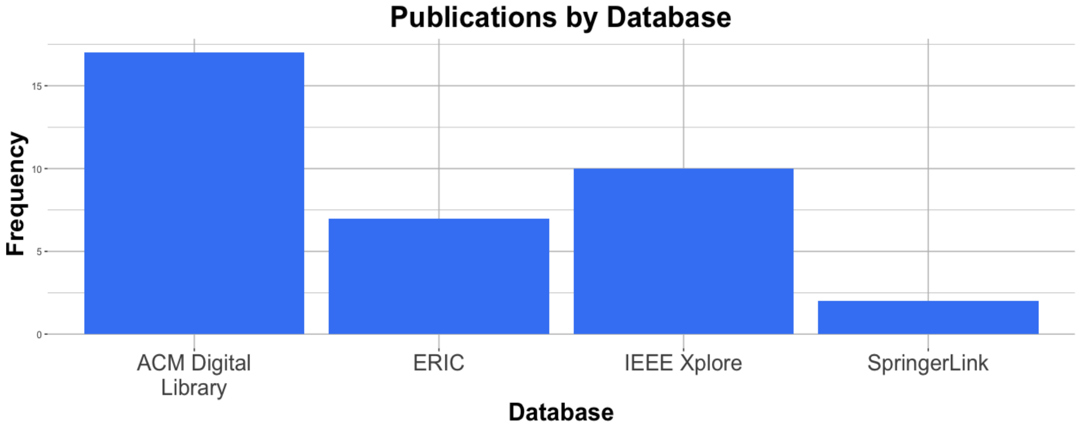
The Reviewed Publications by the Year in Which They Were Published



Most of the records were published in the ACM digital library (47%, n = 17), followed by IEEE (28%, n = 10), ERIC (19%, n = 7), and SpringerLink (6%, n = 2), as shown in Figure 5. Most of the reviewed studies (n = 24) were published in conference proceedings (67%, n = 24), with only 12 published in journals (33%, n = 12). Specifically, most of the studies reviewed were published in conference proceedings in the field of computing science (61%, n = 22) and education (6%, n = 2), followed by journal venues in the fields of computing science (14%, n = 5) and education (19%, n = 7). Also, most studies were published in the field of computing science (75%, n = 27, with 22 proceedings and 5 journals), followed by education (25%, n = 9, with 2 proceedings and 7 journals).

Figure 5

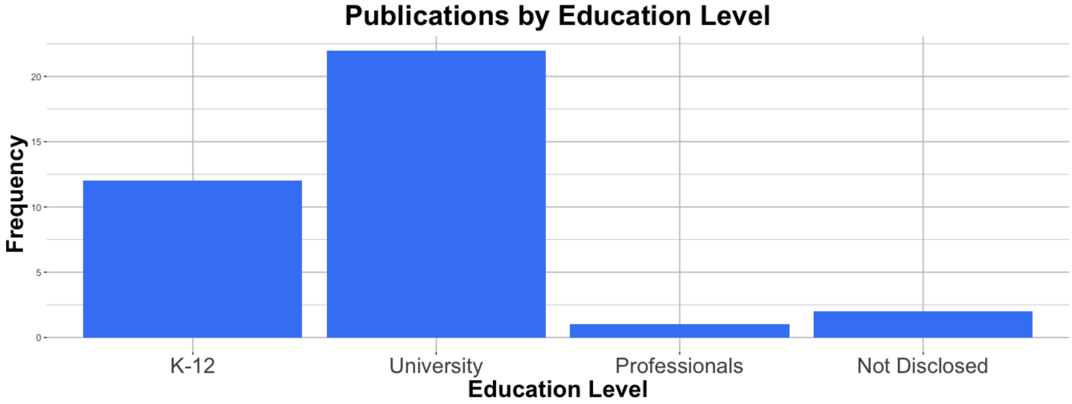
The Publications Reviewed by Searched Database



Most studies were conducted in formal education settings, such as schools, colleges, or universities. Also, most of the studies that we reviewed sampled university or college students (59%, n = 22), followed by K-12 students (32%, n = 12) and professionals (research staff and engineers; 5%, n = 2), as shown in Figure 6. For instance, (VanLehn et al., 2017) conducted an experimental study on both high-school and university students to understand the effectiveness of an interactive tutor offering motivational messages. Five studies sampled a variety of K-12 students: middle school and high school (n = 4) as well as elementary, middle, and high school (n = 1). Finally, one study did not disclose this information.

Figure 6

Publications by Educational Level of the Participants Included in the Reviewed Studies



The 36 records included in this review represented 14 countries across four continents, as shown in Figure 7. The samples included countries from North America (n = 16; USA: n =14, Mexico: n = 2), Europe (n = 10; Norway: n = 4, Germany: n = 1, UK: n = 1, Italy: n = 1, Greece: n = 1, Spain: n = 1, The Netherlands: n = 1), Asia (n = 6; Singapore: n = 2, India: n = 2, China: 1, Taiwan: n = 1), and Australia (n = 1). Five of the publications reviewed failed to specify the sample location. Standen et al. (2020) sampled participants with intellectual disabilities and autism spectrum disorder from three European countries (e.g., The UK, Italy, and Spain) to classify learners' affective states (e.g., engaged, frustrated, or bored).

In terms of author country, most authors were from North America (n = 22; USA: n = 17, Canada: n = 3, Mexico: n = 2), Europe (n = 16; Norway: n = 4, Germany: n = 1, UK: n = 5, Italy:

n = 1, Greece: n = 2, Spain: n = 1, The Netherlands: n = 1, Ireland: n = 1), Asia (n = 6; Singapore: n = 1, India: n = 2, China: 1, Taiwan: n = 1, South Korea: n = 1), and Australia (n = 2). Figure 8 shows the relationship between authors' location and the location of the sample of the studies.

Figure 7

Publications by Continent of the Participants Included in the Reviewed Studies

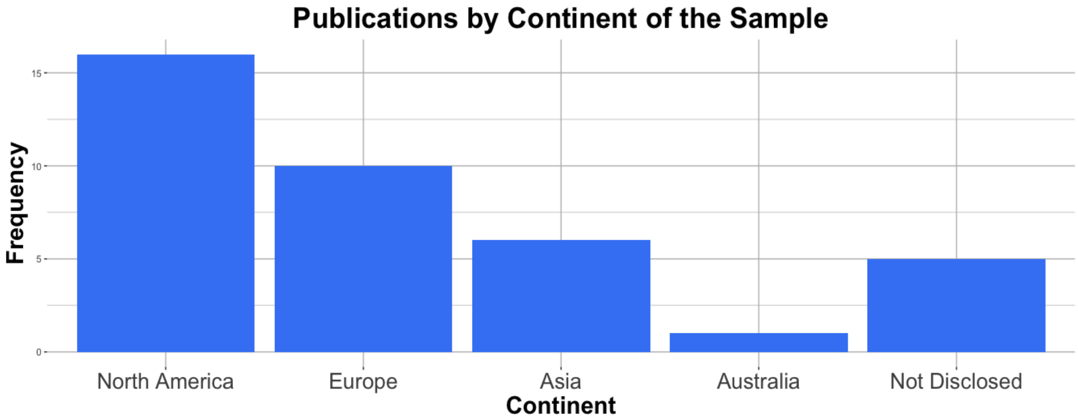
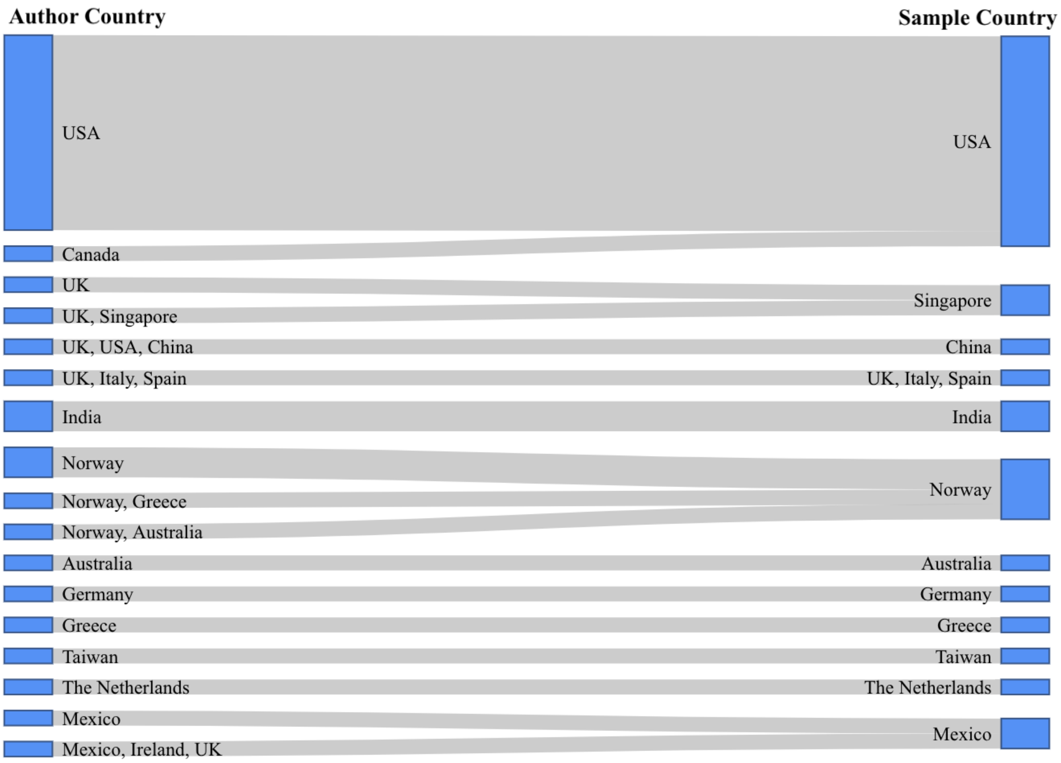


Figure 8

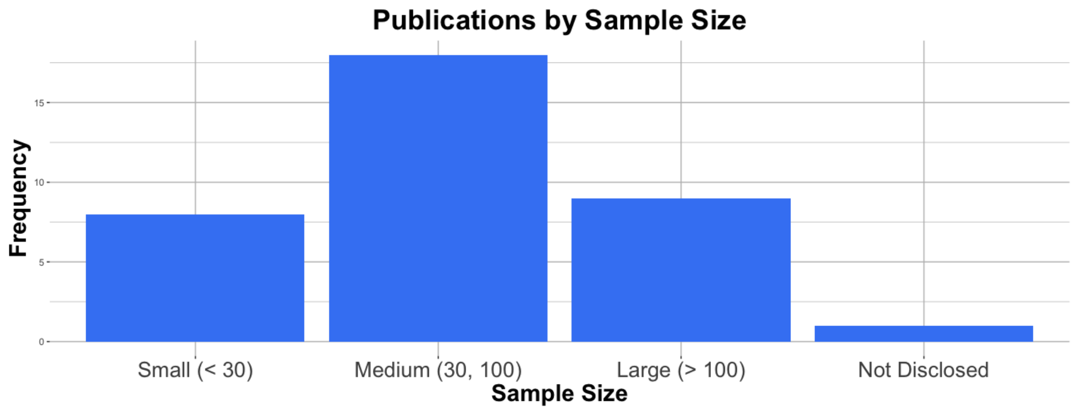
Sankey Chart Linking Author Country to Sample Country



As shown in Figure 9, most of the reviewed publications included medium sample sizes between 30 and 100 participants (n = 18), followed by large sample sizes exceeding 100 participants (n = 9), followed by small sample sizes with less than 30 participants (n = 8). One study did not disclose its sample size.

Figure 9

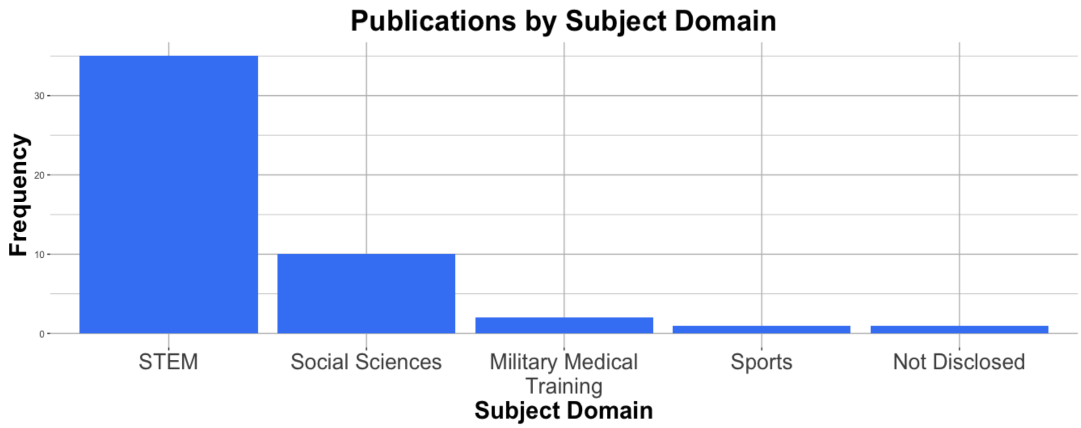
Publications by Sample Size Included in the Reviewed Studies



Regarding the subject domain explored in the reviewed studies, most studies (71.43%) were conducted in the context of STEM, as shown in Figure 10. Many of the STEM studies involved CS and engineering education (n = 12): programming (n = 10), computer and network security (n = 1), and mechanical engineering (n = 1). This was followed by mathematics (n = 12), physics (n = 5), biology (n = 1), and geography (n = 1). Social sciences accounted for 20.41% of the studies (n = 10). The smallest percentage of studies (6.12%) explored the domains of military medical training (n = 2) and sports (n = 1). One study did not disclose the domain of interest.

Figure 10

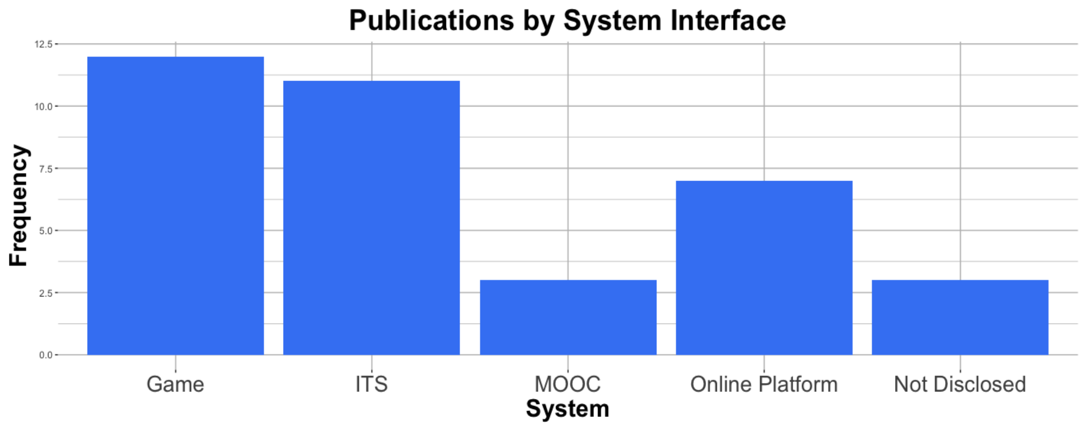
The Reviewed Publications by Subject Domain



Regarding the type of application reviewed, serious games (n = 12; 33.33%) and ITSs (n = 11; 31.1%) were the most employed, as shown in Figure 11. Online platforms were used in n = 7 (19.4%) of the reviewed publications, while MOOCs were used in three (8.3%) of the studies. Three studies did not disclose the type of application used in the analysis.

Figure 11

The Reviewed Publications by the Type of Application



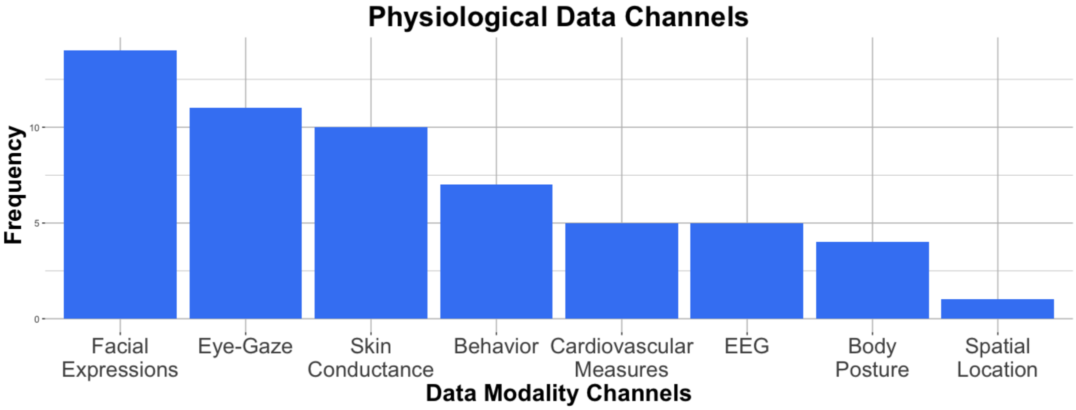
4.2 What are the channels of data employed in the studies reviewed?

4.2.1 Physiological Modality Channels

Several (n = 8) channels of physiological data were identified in the reviewed publications. The most popular channels of data associated with tracking learner activity to infer their affect related to emotions were facial expressions (n = 14), followed by eye-gaze behavior (n = 11), and electrodermal activity (n = 10). This was followed by behavior observation (e.g., upper facial movements, such as eyebrow scrunching; n = 7), EEG measures (n = 5) and cardiovascular measures (n = 5), body posture (n = 4), and the subjects' orientation and location (n = 1), as shown in Figure 12.

Figure 12

The Reviewed Publications by Type of Physiological Data Channels



For each of the physiological channels, the sensors that were used to collect the data, including their brand name, were identified. To collect EEG data, the following devices were used: QUASAR (n = 2) and NeuroSky cap (n = 2). Two studies did not specify the sensor. To collect eye-gaze behavior, the majority of the publications used different devices from the Tobii brand (n = 8; 72.72%): the Tobii X2-30 screen-based eye sensor (n = 2), the wearable Tobii glasses (n = 2), the Tobii 4C (n = 3), and the Tobii X3-120 (n = 1), the most expensive alternative of the screen-mounted sensor. Also, one study reported using a screen-based eye sensor from the brand Eye Tribe and one study did not specify the sensor.

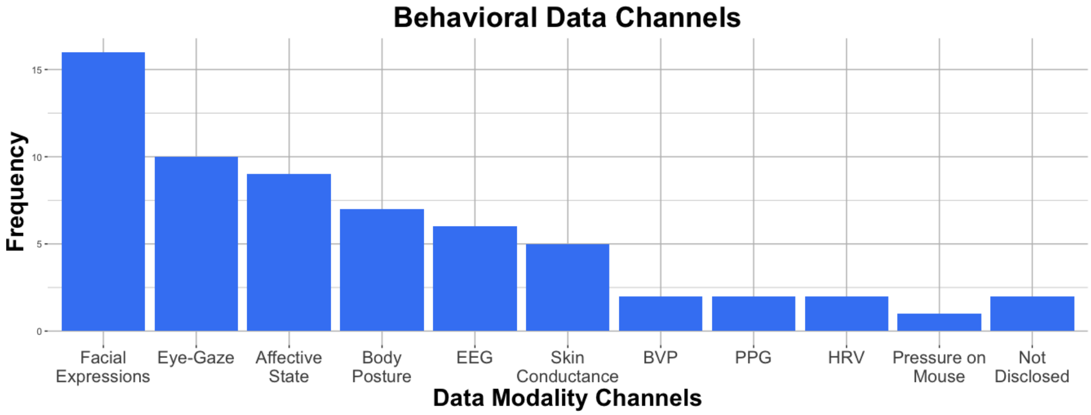
To collect cardiovascular measures, publications reported using the Empatica E4 wristband (n = 1) or an oximeter from an unspecified brand (n = 1). Two studies did not specify the sensor used to collect this channel of data. To collect skin conductance measures, half of the publications reviewed (n = 5) did not specify the brand name of the sensor used, whereas the other half reported using the Empatica E4 wristband (n = 2), the Shimmer bracelet (n = 1), the Q-sensor bracelet (n = 1), and the eSense (n = 1), a device that consists of two nodes attached to the index and middle finger. For spatial location data, a cell phone was used in one of the reviewed publications. Finally, for body posture data (n = 4) and facial expressions (n = 14), no specific (i.e., brand or model) video-based sensor was mentioned.

4.2.2 Behavioral Modality Channels

Behavioral data channels related to the learners' level of involvement in the activity were identified, each sensor was counted individually. From the reviewed publications in descending order of frequency: facial expressions (n = 16), eye-gaze behavior (n = 10), affective state (n = 9), body posture (n = 7), EEG (n = 6), EDA (n = 5), BVP (n = 2), HRV (n = 2), PPG (n = 2), and mouse pressure (n = 1), as shown in Figure 13. To collect behavioral data, the most popular sensor was the Kinect sensor (n = 4), whereas the rest of the publications reviewed did not specify the sensor used (n = 3).

Figure 13

The Reviewed Publications by Type of Behavioral Data Modality Channels

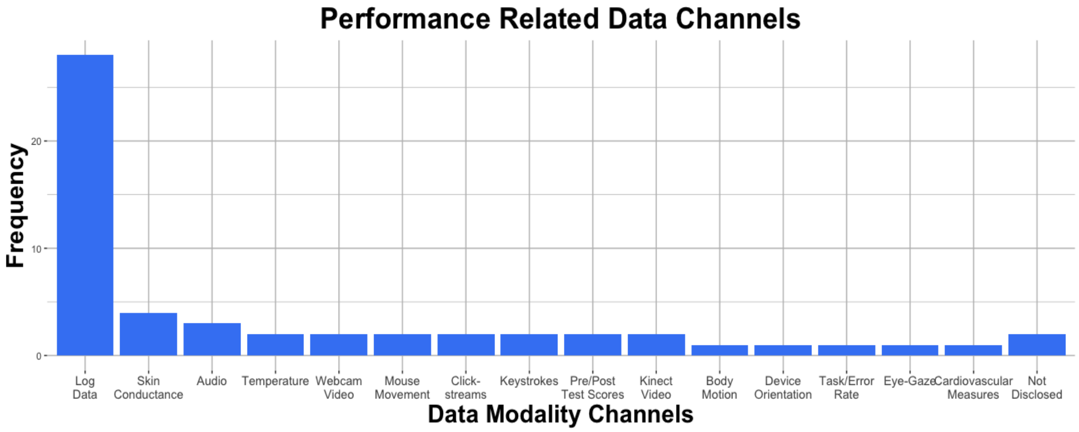


4.2.3 *Performance Related Modality Channels*

Finally, among performance related modality channels, the most popular channel was log data (n = 28; 50%). The rest of the channels identified were not represented as often in the reviewed records: EDA (n = 4), audio (n = 3), mouse movement (n = 2), clickstreams (n = 2), keystrokes (n = 2), pre/posttest performance (n = 2), temperature (n = 2), webcam video (n = 2), cardiovascular measures (n = 1), body motion (n = 1), device orientation (n = 1), task completion and error rate (n = 1), and eye-gaze (n = 1), as shown in Figure 14.

Figure 14

The Reviewed Publications by Type of Performance Related Data Channels



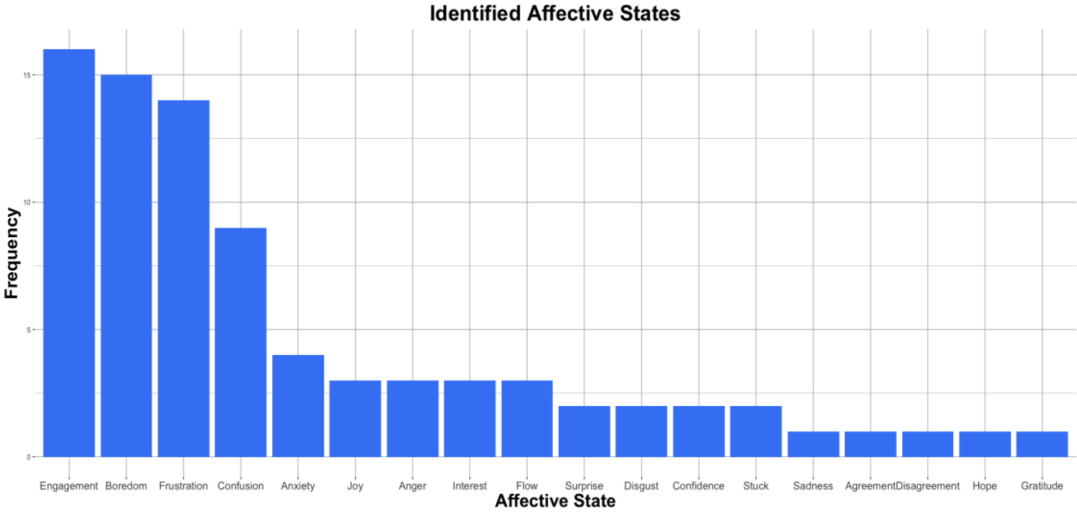
4.3 What are the affective states (classes of emotions) pertaining to the studies reviewed?

Several (n = 18) affective states were identified from the 36 reviewed publications as shown in Figure 15. The most common were engagement (n = 16), boredom (n = 15), frustration (n = 14), and confusion (n = 9). The rest of the recognized affective states seldom repeat themselves: anxiety (n = 4), joy (n = 3), anger (n = 3), interest (n = 3), flow (n = 3), surprise (n = 2), disgust (n = 2), confidence (n = 2), stuck (n = 2), sadness (n = 1), agreement (n = 1), disagreement (n = 1), hope (n = 1), and gratitude (n = 1). Korn and Rees (2019) conducted an experimental study in professionals to identify the affective states the participants experienced when performing tasks in a gamified environment. The identified affective states for this study were: joy, anger, sadness, surprise, fear, contempt, and disgust.

A subset of the affective states of boredom, confusion, concentration, and frustration were commonly (n = 14) found in classification problems and machine learning algorithms were used to distinguish learners' current state among these. For instance, VanLehn et al. (2017) compared several machine learning algorithms using middle school and high school learners' multimodal data when playing a serious game and to classify them among the following affective states: boredom, confusion, delight, engagement, and frustration (Kai et al., 2018).

Figure 15

The Reviewed Publications by Affective States

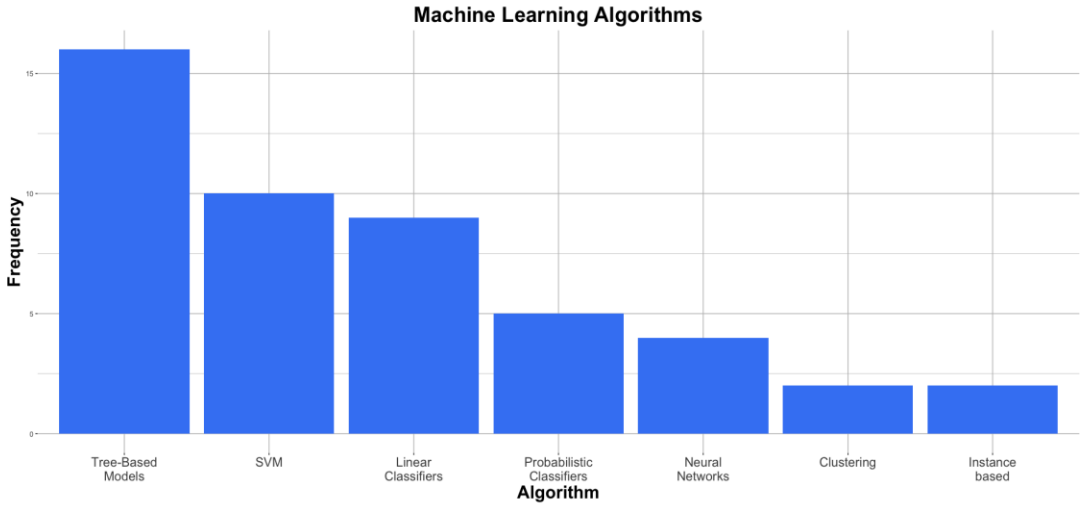


4.4 What are the machine learning algorithms used to detect affect in the studies reviewed?

This review also found that most of the machine learning techniques employed to recognize users' emotional states were supervised learning techniques, specifically classification algorithms. Figure 16 shows that SVMs were the most popular, followed by probabilistic classifiers (e.g., dynamic Bayesian network or DBN, which is a hierarchical probabilistic framework used to classify user concurrent emotions; Naive Bayes), tree-based classifiers (e.g., J48, C4.5), neural networks (e.g., multi-layer perceptron), rule-based (e.g., JRip), instance-based classifiers (e.g., KStar), logistic regression, and other linear classifiers (e.g., step Regression).

Figure 16

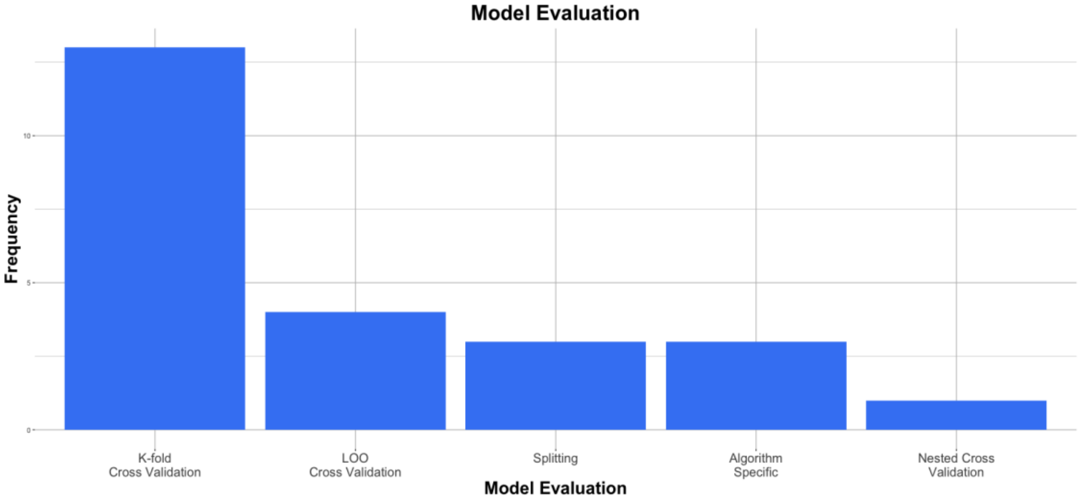
Publications by the Type of Machine Learning Algorithm Used for Classification



From the 25 publications that used at least one machine learning algorithm, all of them reported using a cross-validation technique to evaluate the model, as shown in Figure 17. Most publications (n = 13; 52%) used a k-fold cross-validation technique, followed by Leave-One-Out cross-validation (n = 4), a basic holdout approach (i.e., splitting of the training and test dataset; n = 3), an algorithm-specific approach (n = 3), and nested cross-validation (n = 1).

Figure 17

Publications by Type of Model Evaluation Used on the Machine Learning Algorithms



From the publications that used cross-validation (n = 13), most (n = 7; 53.84%) used 10 folds to evaluate the model, followed by 20 folds (n = 2), 4 folds (n = 4), 6 folds (n = 1), and 1-fold (n = 1). Of the publications that evaluated their model by splitting the training and testing datasets, (n = 1) used a 75/25 split, (n = 1) used a 65/35 split, and (n = 1) was not specific. For the algorithm specific approach, (n = 1) publication used an SVM classifier and adjusted the C parameter (i.e., value of avoiding misclassification at each training example) to 10 and the gamma parameter (i.e., the value to define how far the influence of a single training example reaches) to 0.01. Another (n = 1) used K-Nearest Neighbors and adjusted the number of neighbors (i.e., parameter of nodes to include in the classification voting process) to 10, and, finally, the publication (n = 1) using Random Forests adjusted the number of trees (i.e., number of decision trees to build before averaging the decisions).

Chapter 5. Discussion

5.1 What are the key characteristics of the studies reviewed?

The findings revealed an increase in the number of studies published in recent years. One reason for this result could be linked to the emergence of new, low-cost, and non-invasive wearable sensors becoming available to the public. For example, Empatica Inc. released their first wearable wristband to detect physiological data around 2015 (Comstock, 2015). Similarly, the more popular models of Tobii eye-trackers mentioned in 70% of the reviewed publications that collected eye-gaze behavior were released around 2017. Thus, the increase of publications could be tied to the increase of availability of sensors for research purposes. The increase of publications could also be associated with the publication of the second iteration of the BROMP observation method in 2015 (Ocumpaugh, 2015).

In terms of research design, none of the reviewed publications employed a qualitative methodology. Most of the studies (61.11%) used an experimental approach to their analyses, which is expected because of the quantitative nature of the usage of sensors and the convenience of access to large datasets. However, qualitative research could provide deeper insights into learner affect. Thus, future studies could explore more varied research designs that include interviews, observations, or think-aloud protocols.

The findings revealed that twice as many publications were reported in conference proceedings than in journal venues. Consistent with venue bias (Alshareef et al., 2019), results show that three times more studies were published in the field of computing science than in the field of education. Moreover, twice as many studies across the fields of computing science and education were published in refereed conference proceedings than in journal venues. This result is reasonable, given that some of these publications require advanced machine learning techniques employed in the field of computing science, a fast-moving field, and, increasingly, in the field of education. Although the chosen databases were selected to balance the fields of computing science and education where this research is usually published, three fourths of the reviewed publications were found in the ACM Digital Library and IEEE Xplore databases that mainly focus on publishing computing science venues. Around 90% of the reviewed publications used at least one sensor and 40.90% employed an experimental design to test the hypothesis that sensors could enhance learners' experiences and outcomes. This may be attributable to much of this work requiring the processing of sensor signals. We may see a shift as detectors become commercially available. The choice of an experimental design is likely related to the sample sizes reported in the reviewed publications where 75% of studies collected data from a medium (i.e., between 30 to 100 subjects) to large (i.e., over 100 subjects) of people.

Around two-thirds of the reviewed studies sampled university students. One of the reasons to support this result could be that research consent in younger learners is more difficult to achieve because parents and guardians must be informed and in agreement with the study. Another reason could be that some of the sensors mentioned required placement on the learners' body or limited movement to decrease noise in the data. Moreover, many of the studies (47.2%) were conducted in laboratory settings. Thus, they have limited generalizability to other learning environments, such as classrooms or informal learning settings. Only 38.8% of the studies were conducted in a school laboratory (e.g., computer laboratory) where the data was noisy (Bosch et al., 2015; Kai et al., 2015).

The review found that most studies (75%) employed a limited sample size, especially when exploring particular affective states that occur infrequently (Bosch et al., 2015; Kai et al., 2015). This lack of variability may be due to the variety in the types of learning platforms, which may affect the generalizability of the findings of this review. For example, 33.33% were serious games, 25% were ITSs, 19.44% were online platforms (e.g., Scratch, DuoLingo, etc.), 11.11% were MOOC-based, and 11.11% were not specified. For instance, game-based systems may be conducive to engagement but hinder other affective states. Thus, it would be helpful to conduct more studies sampling large populations, and a variety of system types.

In terms of the sample location, almost half (44%) of the reviewed publications were located in North America. This could be explained by the major investment in sensor-based research across the continent, specifically the United States. In terms of education-based research, Silicon Valley in California (Min et al., 2020) had an increase in investment for the creation of sensors. According to the (Market Research Future, 2018) the North American sensor market has been on the rise since 2016 due to advances in technology and it is expected to double in revenue by 2023. Another explanation for this result could be that the major developers of sensor-based research are located in North America, such as Siemens AG, Honeywell International Inc., General Electric, etc. As most studies sampled students from North America and mostly from classrooms, more research needs to be conducted to explore affect detectors with diverse populations in terms of age, gender, race, and geographic distribution.

Finally, when analyzing the subject domains explored in the reviewed publications, results showed that most studies reviewed focused on STEM contexts (71%). Specifically, there is an interest in exploring how learners react to programming courses which are deemed as difficult and stressful (Demir, 2022). Similarly, serious games and ITSs were the most popular learning applications. These application types are most commonly used in STEM educational contexts. Some examples are Physics Playground (Shute et al., 2013), which is a serious game that teaches fundamental principles of Newtonian physics, and JavaTutor (O'Brien et al., 2014), which is an

ITS that enables learners to visualize their code, execute example problems, and receive reviews from freelance tutors.

5.2 What are the channels of data employed in the studies reviewed?

Results revealed that facial expression was the most common physiological and behavioral data modality channel, being mentioned in a third of the reviewed publications. This result is not surprising because facial expressions are easy to collect and used to explain the affective state of the learner. Also, individuals are more likely to understand non-verbal communication through facial expressions (Buck et al., 1969). Similarly, facial expressions of emotion have been found to exhibit features in common across different cultures (Cowen et al., 2021). Another reason for this result could be that the collection of facial expressions is one of the least disruptive and intrusive data modality channels. Researchers can understand the emotion a learner is experiencing through observation of video streams of the face and computer vision algorithms can vectorize frames of these video streams and identify if the learner expressed an emotion. This identified emotion can then be correlated with the learner's interaction with the learning system. Through the use of video cameras and webcams, learners can be monitored without being physically connected to a sensor or device. This could also explain why eye-gaze behavior was the second most common physiological data channel. However, there are ethical implications of using even the most unobtrusive sensors. For example, using a wristband sensor (e.g., Fitbit or Apple Watch) requires the learner to input their personal information, such as height, weight, and date of birth. This sensitive information can have a negative effect in younger participants as they might see themselves as too heavy or too tall. On the other hand, unsafe storage of sensitive data can result in participant [missing word]. Consequently, additional safeguards and informed consent of the storage, coding, and sharing of the data are necessary.

Finally, log data was the most frequent performance-related data-modality channel. This result is also reasonable because data collected from any type of computer-based systems such as serious games, ITSs, or MOOCs can provide feedback to the researcher about the type of learning experience. This log data can explain if learners are mastering the skill to be learned or if they are struggling to learn it through the number of hints needed or incorrect attempts.

5.3 What are the affective states (classes of emotions) pertaining to the studies reviewed?

The findings revealed that the most commonly explored affective states were engagement, boredom, frustration, and confusion. These states were usually explored together. These results are not surprising given that the literature has identified these affective states as cognitive-affective states (Baker et al., 2010; D'Mello & Graesser, 2011). These states are divided into positive and negative (D'Mello & Graesser, 2011) and have been shown to be influential in learning experiences. This supports Russell's (1980) theory that levels of arousal and emotions have a positive or negative valence and Kort et al.'s (2001) view that emotions are influential and necessary to the learning cycle. For example, negative cognitive-affective states (e.g., frustration) can affect a learning experience positively because they can motivate students to complete challenges (Baker et al., 2021) but they can also affect learning negatively because they can lead to attrition (Vinker & Rubinstein, 2022) and to students feeling unsupported.

To understand learners' affective state through physiological signals, most of the reviewed publications used at least one sensor. These physiological signals can offer researchers more reliable feedback given that these signals originate directly from the sympathetic system and, thus, are more difficult to fabricate, in contrast to behavior that can be controlled such as facial expressions and body posture (Gazzaniga & Smylie, 1990). Likewise, sensors can quantify physiological signals without the added bias from cultural aspects that influence emotion, posited in Hofstede's (2011) Theory of Cultural Dimensions, such as a person's background and how social behaviors like gender roles influence their behavior.

Some of the reviewed publications that used physiological signals to classify learners' affective states (42.85%) were focused on measuring participant engagement in the activity. For example, Lee-Cultura et al. (2020) conducted a study in which middle-school learners played different versions of a serious game in English grammar and math. The authors measured the level of engagement using physiological data channels, such as electrodermal activity and skin temperature collected from an Empatica E4 wristband. Results showed that students felt more represented in the game which led to higher engagement.

Several observation methods were used to classify the learners' affective states. In the BROMP coding protocol, certified observers assign one of the affective labels (i.e., boredom, confusion, delight, engagement, frustration, and surprise) to students. The BROMP protocol was a recurring method in the reviewed publications that were interested in identifying the most common cognitive-affective states (i.e., engagement, boredom, frustration, and confusion). Other reviewed publications employed other types of observation methods to classify the students' affective states, such as surveys, post-tests, and questionnaires. The holistic nature of the observation method (e.g., BROMP) can lead to incomplete data where the coder lacks confidence in the label they assign or misunderstands the learner's behavior. Other observation methods did not have reliability measures as some of the publications used surveys to get feedback in the activity. Although the BROMP observation method is featured repeatedly, there is no agreed upon method to classify learners' affective state through observation. Future studies could compare the effect of the observation method on participants' affective and behavioral states.

The results from the reviewed studies indicate that sensor data can identify different affective states related with achievement, such as engagement, frustration, and boredom (Standen et al., 2020). The absence of boredom was the state most strongly linked to achievement (Standen et al., 2020) because boredom has been found to be a barrier for learning (Chen, 1998). In their experimental study, Standen and colleagues (2020) found significantly more engagement and less boredom in intervention sessions than in control sessions; however, they did not find any significant differences in achievement based on the detection of frustration and engagement. Conversely, Subburaj et al. (2020) conducted a quasi-experimental study sampling undergraduate student triads. The authors used eye-gaze patterns and identified collective gaze agreement (i.e., looking at a similar location for a similar amount of time). Their findings showed that groups with similar eye-gaze sequences reported higher scores when playing an educational game compared to groups that struggled to agree as a team.

The studies rarely showed a relationship between affective state and academic performance (5.55%). Thus, there is an opportunity to develop and validate models that explore this relationship. For example, Kort et al. (2001) suggests that affective states can act as motivation to either continue or abandon a learning cycle. Contrary to machine learning algorithms that require data to be input, teachers can learn to intuitively detect these emotions after forming a relationship with the learners. Both of the studies that predicted academic performance focused on how eye-gaze patterns and behaviors when interacting with a computer-based system correlated with academic performance (Sharma et al., 2019; The & Mavrikis, 2016).

5.4 What are the machine learning algorithms used to detect affect in the studies reviewed?

Only about 10% of the studies reviewed used an unsupervised machine learning method (e.g., clustering) that did not require labeled data. Most of the studies reviewed (91.66%) used supervised learning methods. The findings revealed that the following supervised learning classification algorithms were commonly used in the literature to recognize user emotional states: rule-based reasoning, support vector machines (Bosch et al., 2015; Jraidi et al., 2013; Pham & Wang, 2016), decision trees (Jraidi et al., 2013). One reason for this result can be that these methods are fast and cheap while still yielding valuable models and work well with small amounts of data. In contrast, other algorithms used for affect recognition use past knowledge regarding the user state, such as hierarchical probabilistic methods (e.g., dynamic Bayesian networks or DBN; Jraidi et al., 2013). This result is aligned with the purpose of these studies, which was to predict the affective state of the learner from unimodal or multimodal data channels. One thing to note is that the data used in training and evaluating these supervised learning models requires labels, which can be difficult and costly to produce. The review reveals that it is difficult to definitively compare different types of affect detectors, due to different data sources and particular techniques used to process the data (e.g., missing data techniques) for each data source (e.g., interaction-based versus video-based, etc.). These types of data with different modalities also display a high level of noise, which makes it more difficult to discover useful patterns in the data. More research needs to be conducted to address the issue of noise in multimodal data.

The results revealed that K-fold cross-validation (CV) was the most commonly used evaluation method. This resampling technique splits the dataset into K - 1 subsets for training and one subset for testing. K-fold CV is important, especially when the datasets are smaller because it may reduce bias. Likewise, this method is not as computationally demanding as other resampling methods, such as Leave-One-Out cross validation, and can inform which model yields the best results to previously unseen data.

Regarding vision-based detectors emerging from the studies reviewed include classroom distractions (e.g., fidgeting, talking with one another, asking questions, leaving the computer, and using a cell phone; (Bosch et al., 2015), lighting conditions, and large imbalances in affective state distributions (Bosch et al., 2015; Kai et al., 2015). For example, in some studies, the participant faces could not be captured in a third of the instances even when modern computer vision techniques were employed (Bosch et al., 2015). For instance, some affective states occur at lower rates (e.g., around 5% in (Bosch et al., 2015; Kai et al., 2015), whereas others occurred at higher rates (e.g., around 80%; (Bosch et al., 2015; Kai et al., 2015). Addressing these issues often requires the application of oversampling or under sampling techniques on the training data to avoid creating models that always predict the majority class to the detriment of affective states that are rare but important for learning (e.g., confusion). Likewise, this could be addressed by creating binary classifiers to distinguish between the dominant class and the other classes to find insights on differences among the unbalanced classes.

There was the lack of consistency in the parameters used across models for detecting affect. Differences in performance accuracy of some classifiers were found when different time windows were used. In one study, affective states such as confusion and behaviors such as off-task behaviors were classified better when larger time windows were used, whereas affective states such as delight were better classified using a smaller time window (Bosch et al., 2015). This result is not surprising because behavioral states tend to unfold over longer periods of time (i.e., flow state when actively performing a task in a serious game), contrary to affective states that tend to be situation specific and can be triggered by unexpected stimuli (i.e., sudden change in level of electrodermal activity after receiving pictorial stimuli (Betella et al., 2014; Törmänen et al., 2021). Additionally, there was a lack of consistency in the classifier and feature selection parameters across the studies reviewed. Thus, indicating that the features that are heavily influential have yet to be identified when creating models to predict affective states. Future studies would benefit from tailoring model parameters to particular classification tasks.

5.5 Limitations and Future Work

A limitation of this review is that, although the process of selecting databases and keywords was thorough, it is likely that relevant publications may have been missed. Additionally, the databases used are mostly focused on education and computing, which may have limited the topics covered in the publications. Likewise, the keywords used to perform the search narrowed the results to learning environments, which can explain why there were few studies using samples of professionals. Indeed, almost two-thirds of the samples were from learners at the university level (59%) and only 5% of these were professional learners. This is a limitation because the findings can be difficult to generalize for younger learners (i.e., K-12). Another limitation of the query was that it guided the results towards more quantitative research methods as the exclusion criteria required the studies to use sensors to measure affective state.

Future research can focus on learners in the K-12 level of education to further understand how affective state can predict academic performance. The outcome of this research could aid to the development of safe (i.e., that younger learners cannot swallow or tangle) and more stable sensors (i.e., that may endure movement without introducing noise to the data). Future research would benefit from exploring different databases with a wider range of topics. Furthermore, the choice of keywords could be altered to find a larger variety of sensors and samples.

Chapter 6. Educational Contributions

6.1 Theoretical Contributions

This literature review enhances our understanding of how affective experiences have been designed and developed, which methods have been used to analyze the data, and which research designs have been employed. The current review also shows that the detection of affective states conducive to learning and engagement needs to consider the duration and context of the respective affective event, together with the nature of the interaction (DeFalco et al., 2018). The review also points to the need to develop and validate affective models and interventions (DeFalco et al., 2018).

6.2 Methodological Contributions

This review also highlights the use of data-driven machine learning techniques in detecting affect. Specifically, most of the reviewed studies that used any type of machine learning algorithm (62.5%) built classifiers to detect affect. Affect detection models used different data channels to yield results. Many of the studies (33%) relied on only physical sensor-based detection, with 67% of the studies relying on interaction-based detection (Kai et al., 2015) To detect affective state, the publications focused on different on different information from the learners. Most of the reviewed studies detected user emotions (77.7%), while the rest of the publications focused on detecting learning patterns (13.8%), achievement (5.5%), and receiving feedback from the learners about the educational tool (2.7%). Also, models that use the temporal dynamics of the evolution of affect or behavior during a learner's interaction with the ITS tend to have better predictive accuracy (Joshi et al., 2019; Jraidi et al., 2013). Additionally, some hierarchical probabilistic machine learning approaches (e.g., DBN; Jraidi et al., 2013) can be used to predict interaction trends concomitantly with subsequent affective states.

The results suggest that more research is needed to ascertain whether blending physical sensor-based observation with interaction-based detectors (i.e., within-environment interaction logs) when modeling affective states may increase the utility of each individual modeling approach and act as a backup when data from one of the modeling approaches is missing, especially as this matter is still not elucidated in the explored literature (Kai et al., 2015). In general, some studies found that using several data modalities combined with environment and individual-difference variables leads to better predictive accuracy of a learner's interaction experience and affective states (Jraidi et al., 2013). These combined data modalities can be predictive of learner profile information that can, in turn, be used to trigger tailored interventions that are conducive to better learning (DeFalco et al., 2018).

The review also highlights that physiological sensing often involves non-expensive and non-invasive biofeedback devices that can be used with a variety of systems and provides quantitative information that may be more objective as compared with self-report data collected from questionnaires. Most studies reviewed did not explore the impact of affect on learning, thus, future studies could focus on identifying the context in which affective states have a positive impact on learning outcomes.

6.3 Practical Contributions

The findings of this review may inform the development of adaptive intelligent interfaces for educational software to support learning experiences. These interfaces can detect learner affective states (e.g., engagement) and behavioral states (e.g., off-task behavior). They can also respond in a suitable manner to the affect detected. Thus, such interfaces may adapt to the student depending on the students' affective state (i.e., not intervene if the student is engaged or provide hints, feedback, or suggestions for the next activity otherwise). Also, one of the studies reviewed used facial recognition to predict whether students will answer questions correctly, so that the ITS can adjust the difficulty level of the next question or proactively provide the learner with hints (Joshi et al., 2019).

More research could also be directed towards creating and maintaining domain-specific education datasets that would improve the performance of machine learning algorithms in predicting affect. By constructing these datasets, new models can be optimized through transfer learning (i.e., when an algorithm is storing knowledge that can be used in a related problem; e.g., the algorithm originally learns to classify apples, then it can use that knowledge to identify oranges; Zhuang et al., 2021) and become more accurate at discerning among several affective states and improve from the binary classifiers. However, there is an ethics concern due to the sensitivity nature of the collected data (e.g., educational setting) that can make the individuals easily identifiable.

Chapter 7. Conclusions

In the fields of human-computer interaction (HCI), cognitive science, psychology, education, neuroscience, and computing science, affect detection is increasing in popularity because it aims to improve learner outcomes through adaptation to the learner's affect. In recent years, several models have been built to detect affect in computer-based learning environments. However, affect detection is difficult because emotions cannot be measured directly and can vary from one individual to another. Using physical sensors to collect data about learners' affective states has several limitations (e.g., confinement to laboratory settings, expensive and invasive sensors that are prone to noisy data). This complicates the replication of studies and generalization of results. Even the less invasive sensors can impede the reproduction of studies in younger participants, where sensors require additional supervision and parental consents.

The current review shows that the detection of affective states conducive to learning and engagement needs to consider the duration and context of the respective affective event, together with the nature of the interaction. At least one sensor was used in most of the publications to understand learners' affective state through physiological signals. Some of the reviewed publications that used physiological signals to classify learners' affective states were focused on measuring participant engagement in the activity.

Results revealed that facial expression was the most common physiological and behavioral data modality channel and log data was the most frequent performance-related data-modality channel. Likewise, it was revealed that the most explored affective states were engagement, boredom, frustration, and confusion, which can be identified through sensor data. These affective states can be found in BROMP-based observation, which was a recurring method in the reviewed publications that were interested in identifying the most common cognitive-affective states. Conversely, the studies rarely showed a relationship between affective state and academic performance, suggesting that using sensors to detect affective state is not always beneficial to learning.

This review also highlights the use of data-driven machine learning techniques in detecting affect. Specifically, most of the reviewed studies that used any type of machine learning algorithm-built classifiers to detect affect through supervised learning methods. These methods are easy to compute, require minor data preparation, and are not costly computationally, while still yielding valuable models and work well with small amounts of data.

The results suggest that more research is needed to ascertain whether blending physical sensor-based with interaction-based detectors (i.e., within-environment interaction logs) when modeling affective states may increase the utility of each individual modeling approach and act as a backup when data from one of the sensing approaches is missing, especially as this matter is still not elucidated in the current literature. Some studies found that combining multimodal data channels, environmental, and individual-difference variables led to better predictive accuracy of a learner's interaction experience and affective states.

This review also points to the need to develop and validate affective models and interventions. Among the reviewed publications, there was no agreed-upon method to detect affect from unimodal or multimodal data. Several studies used observation methods that were not validated, such as surveys, which introduces biased data to the models. Also, most of the publications focused on university-level learners, leaving open questions about how these results could be applied to other types of learners (e.g., military trainees, language learners, young and adult learners, students with intellectual disabilities, etc.).

References

Allen, R. L., & Davis, A. S. (2011). Hawthorne effect. In S. Goldstein & J. A. Naglieri (Eds.), *Encyclopedia of Child Behavior and Development* (pp. 731–732). Springer US. https://doi.org/10.1007/978-0-387-79061-9_1324

Alqahtani, F., Katsigiannis, S., & Ramzan, N. (2021). Using wearable physiological sensors for affect-aware intelligent tutoring systems. *IEEE Sensors Journal*, *21*(3), 3366–3378. https://doi.org/10.1109/JSEN.2020.3023886

Alshareef, A., Alhamid, M., & El Saddik, A. (2019). Academic venue recommendations based on similarity learning of an extended nearby citation network. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2019.2906106

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, Ma. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baker, R. D., Gowda, S., & Wixon, M. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In *The Proceedings of the 5th International Conference on Educational Data Mining* (pp. 126–133).

Baker, R., & Ocumpaugh, J. (2014). *Interaction-based affect detection in educational software*. https://doi.org/10.1093/oxfordhb/9780199942237.013.009

Baker, R. S., Nasiar, N., Ocumpaugh, J. L., Hutt, S., Andres, J. M. A. L., Slater, S., Schofield, M., Moore, A., Paquette, L., Munshi, A., & Biswas, G. (2021). Affect-targeted interviews for understanding student frustration. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, &

V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 52–63). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_5

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

Barron-Estrada, M. L., Zatarain-Cabada, R., & Aispuro-Gallegos, C. G. (2018). Multimodal recognition of emotions with application to mobile learning. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 416–418. https://doi.org/10.1109/ICALT.2018.00104

Betella, A., Zucca, R., Cetnarski, R., Greco, A., Lanatà, A., Mazzei, D., Tognetti, A., Arsiwalla, X. D., Omedas, P., De Rossi, D., & Verschure, P. F. M. J. (2014). Inference of human affective states from psychophysiological measurements extracted under ecologically valid conditions. *Frontiers in Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnins.2014.00286

Bickmore, T. W., & Picard, R. W. (2004). Towards caring machines. *Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems - CHI '04*, 1489. https://doi.org/10.1145/985921.986097

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 379–388. https://doi.org/10.1145/2678025.2701397

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624–634. https://doi.org/10.1016/j.jesp.2009.02.016

Buck, R., Savin, V. J., Miller, R. E., & Caul, W. F. (1969). Nonverbal communication of affect in humans. *Proceedings of the Annual Convention of the American Psychological Association*, *4*(Pt. 1), 367–368.

Burleson, W., & Picard, R. W. (2007). Gender-Specific Approaches to Developing Emotionally Intelligent Learning Companions. *IEEE Intelligent Systems*, *22*(4), 62–69. https://doi.org/10.1109/MIS.2007.69

Buzsáki, G., & Watson, B. O. (2012). Brain rhythms and neural syntax: Implications for efficient coding of cognitive content and neuropsychiatric disease. *Dialogues in Clinical Neuroscience*, *14*(4), 345–367. https://doi.org/10.31887/DCNS.2012.14.4/gbuzsaki

Calvo, R. A., & D’Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

Chatterjee, D., Gavas, R., & Saha, S. K. (2022). Exploring skin conductance features for cross-subject emotion recognition. *2022 IEEE Region 10 Symposium (TENSYMP)*, 1–6. https://doi.org/10.1109/TENSYMP54529.2022.9864492

Chen, A. (1998). Perception of Boredom: Students' Resistance to a Secondary Physical Education Curriculum. *Research in Middle Level Education Quarterly*, 21.

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a New General Self-Efficacy Scale. *Organizational Research Methods*, *4*(1), 62–83. https://doi.org/10.1177/109442810141004

Comstock, J. (2015, January 7). *Empatica crowdfunding Embrace, a wearable for epilepsy*. MobiHealthNews. https://www.mobihealthnews.com/39465/empatica-crowdfunding-embrace-a-wearable-for-epilepsy

Covidence. (2019). *World-class systematic review management—A Cochrane technology platform*. https://www.covidence.org/home/

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, *589*(7841), Article 7841. https://doi.org/10.1038/s41586-020-3037-7

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *J. Educ. Media*, *29*(3), 241–250.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper & Row.

Davidson, R. J., Sherer, K. R., & Goldsmith, H. H. (2009). *Handbook of affective sciences*. Oxford University Press.

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193. https://doi.org/10.1007/s40593-017-0152-1

Demir, F. (2022). The effect of different usage of the educational programming language in programming education on the programming anxiety and achievement. *Education and Information Technologies*, *27*(3), 4171–4194. https://doi.org/10.1007/s10639-021-10750-6

Ding, L., Duan, W., Wang, Y., & Lei, X. (2022). Test-retest reproducibility comparison in resting and the mental task states: A sensor and source-level EEG spectral analysis. *International Journal of Psychophysiology*, *173*, 20–28. https://doi.org/10.1016/j.ijpsycho.2022.01.003

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082–1099. https://doi.org/10.1037/a0032674

D'Mello, S., & Graesser, A. (2009). Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence*, *23*(2), 123–150. https://doi.org/10.1080/08839510802631745

D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, *25*(7). https://doi.org/10.1080/02699931.2011.613668

Dorneich, M. C., Whitlow, S. D., Mathan, S., Ververs, P. M., Erdogmus, D., Adami, A., Pavel, M., & Lan, T. (2007). Supporting real-time cognitive state classification on a mobile individual. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 240–270. https://doi.org/10.1518/155534307X255618

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.*, *6*(3–4), 169–200.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement: Investigator's Guide 2 Parts*. Consulting Psychologists Press.

Gazzaniga, M. S., & Smylie, C. S. (1990). Hemispheric mechanisms controlling voluntary and spontaneous facial expressions. *Journal of Cognitive Neuroscience*, *2*(3), 239–245. https://doi.org/10.1162/jocn.1990.2.3.239

Ghaleb, E., Popa, M., Hortal, E., Asteriadis, S., & Weiss, G. (2018). Towards affect recognition through interactions with learning materials. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 372–379. https://doi.org/10.1109/ICMLA.2018.00062

Gosling, S., Rentfrow, P., & Swann, W. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, *37*, 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. *Proceedings of the 16th International Conference on Multimodal Interaction*, 42–49. https://doi.org/10.1145/2663204.2663264

Griffin, S. M., & Howard, S. (2022). Individual differences in emotion regulation and cardiovascular responding to stress. *Emotion*, *22*, 331–345. https://doi.org/10.1037/emo0001037

Han, J., Zhao, W., Jiang, Q., Oubibi, M., & Hu, X. (2019). Intelligent tutoring system trends 2006-2018: A literature review. *2019 Eighth International Conference on Educational Innovation through Technology (EITT)*, 153–159. https://doi.org/10.1109/EITT.2019.00037

Hao, Y., & Foster, R. (2008). Wireless body sensor networks for health-monitoring applications. *Physiological Measurement*, *29*(11), R27. https://doi.org/10.1088/0967-3334/29/11/R01

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, *2*(1). https://doi.org/10.9707/2307-0919.1014

Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: A machine learning workbench. *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference*, 357–361. https://doi.org/10.1109/ANZIIS.1994.396988

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology*.

Johnson, R., & Christensen, L. (2014). *Educational research quantitative, qualitative, and mixed approaches fifth edition*.

Joshi, A., Allessio, D., Magee, J., Whitehill, J., Arroyo, I., Woolf, B., Sclaroff, S., & Betke, M. (2019). Affect-driven Learning Outcomes Prediction in Intelligent Tutoring Systems. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–5. https://doi.org/10.1109/FG.2019.8756624

Jraidi, I., Chaouachi, M., & Frasson, C. (2013). A dynamic multimodal approach for assessing learners' interaction experience. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 271–278. https://doi.org/10.1145/2522848.2522896

Jraidi, I., & Frasson, C. (2013). Student's uncertainty modeling through a multimodal sensor-based approach. *Journal of Educational Technology & Society*, *16*(1), 219–230.

Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision tree modeling of wheel-spinning and productive persistence in skill builders. *Journal of Educational Data Mining*, *10*(1), 36–71. https://doi.org/10.5281/zenodo.3344810

Kai, S., Paquette, L., Baker, R. S., Bosch, N., D'Mello, S., Ocumpaugh, J., Shute, V., & Ventura, M. (2015). A comparison of video-based and interaction-based affect detectors in Physics Playground. In *International Educational Data Mining Society*. International Educational Data Mining Society. http://eric.ed.gov/?id=ED560544

Karray, F., Alemzadeh, M., Saleh, J. A., & Arab, M. N. (2007). Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, *1*(1), 137–159. https://doi.org/10.21307/ijssis-2017-283

Kleinsmith, A., De Silva, P. R., & Bianchi-Berthouze, N. (2006). Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, *18*(6), 1371–1389. https://doi.org/10.1016/j.intcom.2006.04.003

Korn, O., & Rees, A. (2019). Affective effects of gamification: Using biosignals to measure the effects on working and learning users. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 1–10. https://doi.org/10.1145/3316782.3316783

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. *Proceedings IEEE International Conference on Advanced Learning Technologies*, 43–46. https://doi.org/10.1109/ICALT.2001.943850

Lajoie, S. P., & Naismith, L. (2012). Computer-based learning environments. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 716–718). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_512

Lan, A. S., Botelho, A., Karumbaiah, S., Baker, R. S., & Heffernan, N. (2020). *Accurate and interpretable sensor-free affect detectors via monotonic neural networks*. 3.

Lee-Cultura, S., Sharma, K., Papavlasopoulou, S., Retalis, S., & Giannakos, M. (2020). Using sensing technologies to explain children's self-representation in motion-based educational games. *Proceedings of the Interaction Design and Children Conference*, 541–555. https://doi.org/10.1145/3392063.3394419

Leong, F. H. (2015). Automatic detection of frustration of novice programmers from contextual and keystroke logs. *2015 10th International Conference on Computer Science & Education (ICCSE)*, 373–377. https://doi.org/10.1109/ICCSE.2015.7250273

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 298–305. https://doi.org/10.1109/FG.2011.5771414

Liu, Z., Zhi, R., Hicks, A., & Barnes, T. (2017). Understanding problem solving behavior of 6-8 graders in a debugging game. *Computer Science Education*, *27*(1), 1–29.

Mandal, M. K., & Ambady, N. (2004). Laterality of facial expressions of emotion: Universal and culture-specific influences. *Behavioural Neurology*, *15*(1–2), 23–34. https://doi.org/10.1155/2004/786529

Mangaroska, K., Sharma, K., Gaševic, D., & Giannakos, M. (2020). Multimodal learning analytics to inform learning design: Lessons learned from computing education. *Journal of Learning Analytics*, *7*(3), 79–97.

Market Research Future. (2018). *North America Sensor Market Report—Forecast to 2030 | Mrfr* (p. 70). https://www.marketresearchfuture.com/reports/north-america-sensor-market-5194

Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Put your thinking cap on: Detecting cognitive load using EEG during learning. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 80–89. https://doi.org/10.1145/3027385.3027431

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2020). DeepStealth: Game-Based Learning Stealth Assessment With Deep Neural Networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325. https://doi.org/10.1109/TLT.2019.2922356

Mota, S., & Picard, R. W. (2003, June). Automated posture analysis for detecting learner's interest level. *2003 Conference on Computer Vision and Pattern Recognition Workshop*.

Muñoz, K., Noguez, J., Neri, L., Kevitt, P. M., & Lunney, T. (2016). A computational model of learners achievement emotions using control-value theory. *Journal of Educational Technology & Society*, *19*(2), 42–56.

Nasoz, F., Alvarez, K., Lisetti, C. L., & Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, *6*(1), 4–14. https://doi.org/10.1007/s10111-003-0143-x

Neapolitan, R., & Jiang, X. (2007). Probabilistic methods for financial and marketing informatics. In *Probabilistic Methods for Financial and Marketing Informatics*. https://doi.org/10.1016/B978-0-12-370477-1.X5016-6

Nowlis, H. H., & Nowlis, V. (1956). The description and analysis of mood. *Annals of the New York Academy of Sciences*, *65*(4), 345–355. https://doi.org/10.1111/j.1749-6632.1956.tb49644.x

Nurhudatiana, A., Hiu, A. N., & Ce, W. (2018). Should I use laptop or smartphone? A usability study on an online learning application. *2018 International Conference on Information Management and Technology (ICIMTech)*, 565–570. https://doi.org/10.1109/ICIMTech.2018.8528134

O'Brien, C., Goldman, M., & Miller, R. C. (2014). Java Tutor: Bootstrapping with Python to learn Java. *Proceedings of the First ACM Conference on Learning @ Scale Conference*, 185–186. https://doi.org/10.1145/2556325.2567873

O'Brien, H., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined User Engagement Scale (UES) and new UES Short Form. *International Journal of Human-Computer Studies*, *112*. https://doi.org/10.1016/j.ijhcs.2018.01.004

Ocumpaugh, J. L. (2015). *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.* https://www.academia.edu/10941533/Baker_Rodrigo_Ocumpaugh_Monitoring_Protocol_BROMP_2_0_Technical_and_Training_Manual

Oliveira, P. C. de, Cunha, C. J. C. de A., & Nakayama, M. K. (2016). Learning management systems (LMS) and e-learning management: An integrative review and research agenda.

JISTEM - Journal of Information Systems and Technology Management, *13*, 157–180. https://doi.org/10.4301/S1807-17752016000200001

Paas, F., & Ayres, P. (2014). Cognitive load theory: A broader view on the role of memory in learning and education. *Educational Psychology Review*, *26*(2), 191–195. https://doi.org/10.1007/s10648-014-9263-5

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *J. Clin. Epidemiol.*, *134*, 178–189.

Pardos, Z. A., Baker, R. S. J. D., San Pedro, M. O. C. Z., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, *1*(1), 107–128.

Park, J., Park, Y. H., Kim, J., Cha, J., Kim, S., & Oh, A. (2018). Elicast: Embedding interactive exercises in instructional programming screencasts. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10. https://doi.org/10.1145/3231644.3231657

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). Chapter 2 - The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in Education* (pp. 13–36). Academic Press. https://doi.org/10.1016/B978-012372545-5/50003-4

Peng, J., Lee, K., & Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research - J EDUC RES*, *96*, 3–14. https://doi.org/10.1080/00220670209598786

Pham, P., & Wang, J. (2016). Adaptive review for mobile MOOC learning via implicit physiological signal sensing. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 37–44. https://doi.org/10.1145/2993148.2993197

Picard, R. W. (1997). *Affective computing*. MIT Press.

Psaltis, A., Apostolakis, K., Dimitropoulos, K., & Daras, P. (2017). Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Computational Intelligence and AI in Games*, *10*, 1–1. https://doi.org/10.1109/TCIAIG.2017.2743341

Rajendran, R., Iyer, S., & Murthy, S. (2019). Personalized affective feedback to address students' frustration in ITS. *IEEE Transactions on Learning Technologies*, *12*(1), 87–97. https://doi.org/10.1109/TLT.2018.2807447

Rajendran, R., & Muralidharan, A. (2013). Impact of Mindspark's adaptive logic on student learning. *2013 IEEE Fifth International Conference on Technology for Education (T4e 2013)*, 119–122. https://doi.org/10.1109/T4E.2013.36

Robison, J., McQuiggan, S., & Lester, J. (2009, September). Evaluating the consequences of affective feedback in intelligent tutoring systems. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*.

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

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*, 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Saxena, P., Dabas, S., Saxena, D., Ramachandran, N., & Ahamed, S. I. (2020). Reconstructing compound affective states using physiological sensor data. *2020 IEEE 44th Annual*

Computers, Software, and Applications Conference (COMPSAC), 1241–1249. https://doi.org/10.1109/COMPSAC48688.2020.00-86

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology*, *44*, 229–237. https://doi.org/10.1037/h0055778

Sharma, K., Papavlasopoulou, S., & Giannakos, M. (2019). Coding games and robots to enhance computational thinking: How collaboration and engagement moderate children’s attitudes? *International Journal of Child-Computer Interaction*, *21*, 65–76. https://doi.org/10.1016/j.ijcci.2019.04.004

Sharma, K., Papavlasopoulou, S., Lee-Cultura, S., & Giannakos, M. (2021). Information flow and children’s emotions during collaborative coding: A causal analysis. *Interaction Design and Children*, 350–362. https://doi.org/10.1145/3459990.3460731

Shim, K.-S., Kim, Y., Sohn, I., Lee, E., Bae, K., & Lee, W. (2022). Design and validation of quantum key management system for construction of KREONET quantum cryptography communication. *Journal of Web Engineering*. https://doi.org/10.13052/jwe1540-9589.2151

Shute, V. J., & Psotka, J. (1994). *Intelligent Tutoring Systems: Past, Present, and Future.:* Defense Technical Information Center. https://doi.org/10.21236/ADA280011

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton’s Playground. *The Journal of Educational Research*, *106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

Sinha, A., Gavas, R., Chatterjee, D., Das, R., & Sinharay, A. (2015). Dynamic assessment of learners’ mental state for an improved learning experience. *2015 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE.2015.7344121

Sofi Dinesh, Rejikumar, G., & Sisodia, G. S. (2021). An empirical investigation into carpooling behaviour for sustainability. *Transportation Research Part F: Traffic Psychology and Behaviour*, *77*, 181–196. https://doi.org/10.1016/j.trf.2021.01.005

Sottilare, R. A., & Proctor, M. (2012). Passively classifying student mood and performance within intelligent tutors. *Journal of Educational Technology & Society*, *15*(2), 101–114.

Srivastava, N., Newn, J., & Velloso, E. (2018). Combining low and mid-level gaze features for desktop activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(4), 189:1-189:27. https://doi.org/10.1145/3287067

Standen, P. J., Brown, D. J., Taheri, M., Galvez Trigo, M. J., Boulton, H., Burton, A., Hallewell, M. J., Lathe, J. G., Shopland, N., Blanco Gonzalez, M. A., Kwiatkowska, G. M., Milli, E., Cobello, S., Mazzucato, A., Traversi, M., & Hortal, E. (2020). An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *British Journal of Educational Technology*, *51*(5), 1748–1765. https://doi.org/10.1111/bjet.13010

Subburaj, S. K., Stewart, A. E. B., Ramesh Rao, A., & D'Mello, S. K. (2020). Multimodal, multiparty modeling of collaborative problem solving performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 423–432). Association for Computing Machinery. http://doi.org/10.1145/3382507.3418877

Terras, M. M., & Ramsay, J. (2015). Massive open online courses (MOOCs): Insights and challenges from a psychological perspective. *British Journal of Educational Technology*, *46*(3), 472–487. https://doi.org/10.1111/bjet.12274

The, B., & Mavrikis, M. (2016). A study on eye fixation patterns of students in higher education using an online learning system. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 408–416. https://doi.org/10.1145/2883851.2883871

Törmänen, T., Järvenoja, H., & Mänty, K. (2021). All for one and one for all – How are students' affective states and group-level emotion regulation interconnected in collaborative learning? *International Journal of Educational Research*, *109*, 101861. https://doi.org/10.1016/j.ijer.2021.101861

Vail, A. K., Grafsgaard, J. F., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2016). Gender differences in facial expressions of affect during learning. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 65–73. https://doi.org/10.1145/2930238.2930257

VanLehn, K., Zhang, L., Burleson, W., Girard, S., & Hidago-Pontet, Y. (2017). Can a non-cognitive learning companion increase the effectiveness of a meta-cognitive learning strategy? *IEEE Transactions on Learning Technologies*, *10*(3), 277–289. https://doi.org/10.1109/TLT.2016.2594775

Veliyath, N., De, P., Allen, A. A., Hodges, C. B., & Mitra, A. (2019). Modeling students' attention in the classroom using eyetrackers. *Proceedings of the 2019 ACM Southeast Conference*, 2–9. https://doi.org/10.1145/3299815.3314424

Villon, O., & Lisetti, C. (2006, September). A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*.

Vinker, E., & Rubinstein, A. (2022). Mining code submissions to elucidate disengagement in a Computer Science MOOC. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 142–151. https://doi.org/10.1145/3506860.3506877

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, *7*, 225–240. https://doi.org/10.1162/105474698565686

Wu, C.-H., Tzeng, Y.-L., & Huang, Y.-M. (2020). Measuring performance in leaning process of digital game-based learning and static E-learning. *Educational Technology Research and Development*, *68*(5), 2215–2237. https://doi.org/10.1007/s11423-020-09765-6

Xiao, X., & Wang, J. (2015). Towards attentive, bi-directional MOOC learning on mobile devices. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 163–170. https://doi.org/10.1145/2818346.2820754

Xiao, X., & Wang, J. (2016). Context and cognitive state triggered interventions for mobile MOOC learning. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 378–385. https://doi.org/10.1145/2993148.2993177

Yang, T.-Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active learning for student affect detection. *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019*, 208–217.

Yu, C.-H., Wu, J., & Liu, A.-C. (2019). Predicting learning outcomes with MOOC clickstreams. *Education Sciences*, *9*(2), 104. https://doi.org/10.3390/educsci9020104

Yue, J., Tian, F., Chao, K.-M., Shah, N., Li, L., Chen, Y., & Zheng, Q. (2019). Recognizing multidimensional engagement of e-learners based on multi-channel data in e-learning environment. *IEEE Access*, *7*, 149554–149567. https://doi.org/10.1109/ACCESS.2019.2947091

Zhu, M., Sari, A. R., & Lee, M. M. (2020). A comprehensive systematic review of MOOC research: Research techniques, topics, and trends from 2009 to 2019. *Educational Technology Research and Development*, *68*(4), 1685–1710. https://doi.org/10.1007/s11423-020-09798-x

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Appendix A: Detailed Summaries of the Reviewed Studies

Barron-Estrada et al. (2018) conducted an experimental study sampling n = 10 (3 females) university students enrolled at the Culiacan Institute of Technology in Mexico. The study presented the initial implementation of a system to detect learners' emotion using mobile devices, specifically boredom and engagement. The proposed system does not require the use of invasive and expensive sensors, as it only requires a mobile device that is able to run the EmoData application, a software to collect information from the user using the device components, such as the microphone, accelerometer, and gyroscope. The authors also developed an affective database with the help of the EmoData application.

The students interacted with the popular language learning app, Duolingo, while using the mobile device that had the EmoData application installed. The authors describe the Duolingo application as ideal for their study because they could collect audio recordings for English pronunciation exercises for Spanish speakers.

Each of the channels (audio, position, and movement) was classified unimodally (i.e., each gets an emotion, resulting in three emotions), then these emotions are merged through a "fuzzy inference system". The fuzzy system uses rules to determine the emotion (Engagement, Boredom, or None). The results of only using the accelerometer and gyroscope to recognize posture without using the EmoData app information yielded a 73% of accuracy, whereas the results of only using the audio recognition yielded a 50% of accuracy.

One limitation mentioned by the authors was the small sample, which affected the accuracy of their system.

Bosch et al. (2015) conducted an online quasi-experimental (qualitative pre-/post-test testing physics knowledge and skills) classroom study that developed and validated face-based detectors of learning-centered affect (boredom, confusion, delight, engaged concentration, and frustration) and of student behaviors: *on task* (i.e., when looking at their own computer), *on-task conversation* (i.e., when conversing with other students about the task), and *off-task* (i.e., in other situations, such as using a cell phone) in a physics game, *Physics Playground*, which was designed to teach principles of Newtonian physics. This study uses multimodal data sources. It showed that learning-centered affective states can be identified from naturalistic facial expressions and from body movements in a school context. In this study, facial expression recognition was applied to video data collected via a webcam.

This study sampled n = 137 (80 female) Grade 8-9 US public-school students. Students were tested in groups of 20. They played Physics Playground during regular 55-minute class periods over four days. Live field observations of learners' on-task versus off-task behaviors were collected during gameplay by two Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP 2.0; Baker & Ocumpaugh, 2014) certified observers. Usually, BROMP is implemented using the Android app, Human Affect Recording Tool (HART; Baker et al., 2012). The observations constituted the ground truth affect and behavior annotations used in training automated detectors for both types of detectors to predict learners' affective states and off-task behaviors. There were 1767 successful observations of affective stats and 1899 observations of on-task/off-task behavior.

For the video-based detection, 78 facial features were extracted using the FACET version of the Computer Expression Recognition Toolbox (CERT) software (Littlewort et al., 2011). FACET is used to automatically detect 19 Action Units (AUs; Ekman & Friesen, 1978) as well as orientation and position of the face, which constitute labels for specific activations of facial muscles (e.g., lowered brow). As before, the AU-labeled data was temporally aligned in small time windows with observations of affect to create features. Features were created by aggregating (i.e., maximum, median, and standard deviation) FACET values (i.e., AUs, face orientation, and face position) in a window of time leading up to each observation. For video-based models, body movement features were also extracted by measuring the proportion of pixels in each video frame that differed from a continuously updated estimate of the background image generated from the four previous frames. One detector model was built for the overall five-way affect discrimination (bored, confused, delighted, engaged, and frustrated). Separate binary (i.e., two-class) detector models were built for each affective state, where the affective state (e.g., engaged concentration) was discriminated from the combination of the rest of the instances (e.g., frustrated, bored, delighted, and confused combined, referred to as "all other"). This enables the parameters (e.g., window size, features used) to be optimized for that particular affective state. Behaviors were also grouped into two classes: off-task behaviors versus the not off-task behaviors (i.e., combined *on-task behaviors* and *on-task conversation behaviors* related to the game). Fourteen different classifiers were used to build models for seven discriminations: overall five-way classifier, five affective state models, and off task vs. on task model) and they included support vector machines (SVMs), C4.5 trees, and Bayesian classifiers from the WEKA machine learning tool (Holmes et al., 1994).

Model evaluation was conducted using a 10-fold nested cross-validation process on student data and the performance metric used was A' (i.e., the probability of correctly determining if an instance belongs to a certain affective state) or the area under the receiver operating characteristic (ROC) curve. The results revealed that classification was successful, exceeding chance for both data modalities. The classification results showed that performance exceeded chance for off-task behavior (AUC = .816) and each affective state: boredom (.610), confusion (.649), delight (.867), engagement (.679), and frustration (.631) as well as a five-way overall classification of affect (.655).

One limitation of this study is that the link between affect and learning was not explored.

Burleson & Picard (2007) conducted an experimental study to investigate students' levels of engagement when playing Towers of Hanoi, an educational game to learn mathematics. For this game, multimodal non-verbal data was used to develop a game companion that interacts with the learner based on their sensor information in real-time. The authors address the importance of the social bond between instructors and middle schoolers, as it can significantly predict their learning performance. The researchers collected facial expressions using a camera, electrodermal activity, posture using a sensor driven chair, and logs from the game.

The study collected data from n = 76 middle school (age range: 11 to 13 years) students from three semi-rural schools in western United States. The learners were randomly assigned to one of the two strategies: (1) sensor driven non-verbal mirroring, where the sensors were used to mimic the participants behavior, and (2) pre-recorded interactions generated from a previous pilot study. Additionally, the participants were randomly assigned to two interventions: (1) affective support intervention targeting the learners' current emotion, and (2) task support intervention to help the student struggling in the task. The intervention started with a pre-test to assess the learners' self-theories of intelligence and goal mastery orientation. Then, the digital character presented the activity and, after four minutes, the character started interacting with the learner depending on the strategy and intervention assigned. During the communication, the companion asked the student to report their frustration level on a 7-point scale. Afterwards, the students answered post-activity questions about the activity (e.g., asking how much of the time interacting with the game they felt frustrated, etc.). Then, a 1.5 calming video was shown, as a neutral affect inducement. Finally, the students answered the post-test, identical to the pre-test, and the Working Alliance Inventory, to assess their impression on the character (Bickmore & Picard, 2004).

The researchers found differences with the social bond girls and boys developed. Results from ANOVA analyses demonstrated that the intervention had opposing effects for boys and girls, as boys responded more positively in the task support condition than in the affect support condition, while girls presented an opposite relationship from boys in both interventions. Moreover, girls were more motivated by frustration in either intervention, whereas boys were more perseverant when they received sensor driven non-verbal interactions. Finally, girls were less frustrated than boys at the end of the intervention.

DeFalco et al. (2018) conducted three experimental studies to detect undergraduate students' frustration in a modified serious game for military training, the TC3Sim combat medic care training course that provides training for first responders. Authors used sensor-based and interaction-based sensor-free affect detectors that were developed as part of their three-year project. The study aims to integrate and unify affect detection, validation, and intervention, and to investigate whether linking different types of affective interventions to detectors improves students' learning outcomes. The results of the first study revealed a negative correlation between frustration and learning gains when participants engaged in the TC3Sim game and were further used to inform the development of the affect detectors. The second study integrated the interaction-based detectors into the TC3Sim game-based simulation through affect recognition functionalities provided by the GIFT framework for intelligent tutoring. The results of the second study revealed that motivational intervention feedback messages designed to address student frustration (i.e., self-efficacy enhancing feedback interventions based on interaction-based affect detectors) yielded significantly greater learning gains than control conditions that did not include motivational feedback messages. The third experiment compared feedback messages triggered by the sensor-based affect detectors (i.e., the Kinect-based detectors) and the interaction-based sensor-free detectors. The findings of the third study showed that the Kinect-based detectors did not detect participant frustration. Also, there were no differences in learning among (1) interventions triggered by the interaction-based detector, (2) interventions triggered on a fixed schedule, and (3) no interventions at all.

Ghaleb et al. (2018) conducted a correlational study to evaluate a model that detected learners' affective state, such as boredom, engagement, and frustration. Based on the Theory of Flow (Csikszentmihalyi, 1990), which refers to the satisfaction of individuals when they are fully immersed in performing an activity. The researchers designed a serious game where the students answer questions in different formats (i.e., multiple answers, true or false, and fill in the blank) from major topics, such as Mathematics, History, Sports, and Geography. The question database included 800 questions with different levels of difficulty that produce different affective states.

The study collected data from n = 32 (18 females) bachelor (n = 20) or master (n = 12) students from the University of Maastricht in the Netherlands. Each student performed four sessions for each of the four topics. The average duration of the intervention per participant lasted 26 minutes. The researchers collected n = 459 sessions from the students, which were labeled by self-reported affective states of the participants. The logs from the interaction with the game were recorded using xAPI, an event-centered framework for tracking and storing educational data (Santos et al., 2015) covering a larger range of actions.

An SVM classifier with a radial basis function (RBF) kernel was trained using a one-versus-all strategy to classify emotions. For example, the first model was a binary classifier, that distinguished between frustration and a combined class of boredom and engagement. The second and third models were created using the same strategy. Finally, the three models were combined into a hidden layer to create a more robust model. Two settings of cross validation were applied to the model: (1) cross-subject validation, to test generality across students with different profiles, and (2) subject-based validation, to enable adaptive learning according to the learners' personality.

The cross-subject model applied leave-one-out-cross-validation (LOOCV) and 10-fold cross validation yielded a precision of 66%. As expected, the model was more accurate at detecting engagement and frustration because boredom was the least reported affective state. Afterwards, to improve model precision, a binary classifier was built to detect between engagement and not engagement, which yielded a precision of 75%. The subject-based model reported a 74% precision and was able to detect sudden changes in affective states.

In sum, the proposed models reported a high classification accuracy. However, the results from the subject-based model highlight the relevance of a learners' affective state and interactive features to offer personalized learning activities. For example, results show a higher detection accuracy of engagement than non-engagement for engineering students with an exception in sport. For psychology students, in all topics, the detection of non-engagement was relatively higher than engagement. Overall, engineering students outperformed psychology students, which might explain their engagement in playing the game. For psychology students, the non-engagement can be due to the perceived challenge and their lower interest towards the game topics. These observations and results are consistent with the Theory of Flow model.

Grafsgaard et al. (2014) conducted a quasi-experimental study to determine how multimodal feature sets are used to predict the overall affective state a student experiences during an entire learning session. Using the JavaTutor platform to visualize Java code, the researchers collected several streams of data, such as recordings from the sessions (i.e., webcam video, Kinect depth video, database logs), audio recording for dialogue analysis from human tutors communicating with the students through the platform, physiological data (i.e., electrodermal activity), and non-verbal behavior (e.g., facial expressions, hand-to-face gestures, and posture). The authors built different classifiers using unimodal, bimodal, and multimodal data and compared them to identify which feature set was the most predictive.

The study collected data from n = 67 (average age 18.5 years) university students in the United States enrolled in an introductory engineering course. Before each session using the JavaTutor platform, the students completed a content-based pretest. After each session, the students completed a posttest, which was identical to the pretest, and a post-session survey that included: (1) the User Engagement Survey (UES) to measure aesthetic appeal, focused attention, novelty, usability, involvement, and endurability in human-computer interactions (HCI; O'Brien et al., 2018) and (2) the NASA-TLX workload survey, which included a question regarding the learner's frustration level.

Three data streams were established to build the models: (1) dialogue, which included dialogue messages and their respective answers from the platform, (2) task, which included database logs from the platform (e.g., compile attempts, running the program, etc.), and (3) nonverbal, which included the student's facial expression, hand to face gestures, and posture. The models were built using linear regression and using leave-one-out cross-validation.

To predict engagement, eight classifiers were built: three used unimodal data, three used combinations of bimodal data, and two used combinations of the three streams of data. Results from the R^2 value show that the trimodal model combining dialogue, nonverbal behavior, and task actions was the best performing model ($R^2 = 0.282$). A similar approach was conducted to build the model to predict frustration. As in the previous engagement detector model, the trimodal model combining dialogue, nonverbal behavior, and task actions outperformed the rest of the models ($R^2 = 0.520$). Finally, to detect normalized learning gain using the pretest and posttest results, a similar approach to the previous models was conducted, thus creating eight models. Just as in the previous models, the trimodal model combining the three data streams outperformed the rest of the models ($R^2 = 0.544$). However, in contrast to the models that predicted affective states, the unimodal model using only the dialogue data stream had significant predictive power ($R^2 = 0.370$).

In sum, multimodal feature sets are the most predictive classifiers when detecting affective state. Dialogue was a significant predictor when detecting learning gain in the three levels of this classifier.

Joshi et al. (2019) conducted a correlational study sampling n = 30 (26 females) college students. The study predicted students' learning outcomes (i.e., correctness of responses to mathematics questions) from facial affect signals (i.e., action unit-based feature representations) based on videos of student interactions with the MathSpring ITS as they begin to solve mathematics problems. It also attempts to predict how early the student learning outcome can be predicted, so appropriate interventions can be provided. The multimodal detector uses signals from a laptop webcam and a GoPro camera placed on the trackpad of the laptop to capture students' facial expressions, a video stream of the screen activity (ITS interface and users mouse interactions with the ITS interface), and mouse movements (location trajectories and clicks). Importantly, the binary classification model built in this study used the temporal dynamics of the evolution of facial behavior during a learner's interaction with the ITS. This study attempts to elucidate the relationship between facial features extracted from a video stream and learning outcomes. One of the contributions of this study is the creation of a facial affect database containing 38 videos of college students' interactions (n = 1596) with the MathSpring ITS, with each interaction lasting for approximately an hour. Each instance used to train the machine learning classifier consists of a video clip of the student solving a problem together with the outcome (solved or not). OpenFace, a facial behavior analysis toolkit (Baltrušaitis et al., 2016), detected 17 Action Units (AUs) from the video stream data. From each frame of all the video streams in the data set, the mean Action Unit Occurrence (AUO) for 18 AU presence and 17 AU intensity values was computed. Additionally, head-pose and eye-gaze vectors were extracted. Finally, a 376-dimensional feature representation was used for a multi-layer perceptron with 2 hidden layers, each with 100 activation nodes, and the Adam optimizer to predict several classes of effort. Model training was performed using the first 1, 5, 10, and 30 seconds, along with the entire length of the input. The baseline model trained and tested on the entire input length yielded a mean accuracy of 0.54 and a mean F1-Score of 0.27. Also, individual one-vs-all binary classifiers were trained to predict all seven effort labels. The results show that model performance for both the multiclass and binary classifiers increases when features are computed from longer temporal sequences. The findings also show that, overall, the baseline models predict more accurately whether a student eventually answered correctly after seeing one or more hint, whether a student performed some action but without having read the problem, and whether a student answered correctly on the first attempt without seeing any hints, in comparison to predicting whether a student did not see any hints but solved the question after one incorrect attempt, whether a student performed some action but did not solve the problem at all, whether a student did not see hints but solved the question after greater than one incorrect attempt, and whether a student skipped the problem with no action.

Jraidi & Frasson (2013) conducted a quasi-experimental study to propose a multimodal sensor-based model to detect learners' uncertainty based on electroencephalogram (EEG) activity that measured cognitive load, affective state, and personal characteristics. The study had two main purposes: (1) to find the behavior trends related to uncertainty and (2) to build a model that predicts uncertainty. Data were collected via three physiological channels: EEG to measure mental concentration, skin conductance (SC) to measure arousal, and blood volume pulse (BVP) to measure valence of arousal (i.e., positive, or negative).

The study collected data from n = 38 (14 females) recruited learners, with an average age of 27.31 years. The experiment was presented upon arrival to the laboratory setting and the students were set up with the sensors. Learners completed the tasks on a problem-solving Intelligent Tutoring System (ITS) while the researchers recorded the response time and the answer for each question. Finally, the students answered the following surveys: (1) demographics survey, (2) self-perceived logical problem-solving skill level, and (3) Big Five Inventory to measure personality dimensions.

An ANOVA analysis was performed to find relationships between learners' uncertainty and affective reactions. Results suggest that when learners are uncertain, they may be more focused on trying to solve a problem, whereas when learners are certain, they may be at ease and not very concentrated.

To predict uncertainty, binary classifier algorithms were compared: Decision Tree, Naïve Bayes, and SVM. WEKA was used to train the classifiers and K-fold cross-validation was used to evaluate the model. The SVM classifier yielded the highest performance (83.25% accuracy).

Jraidi et al. (2013) conducted an experimental study sampling n = 44 learners that interacted with three learning environments solving cognitive tasks in the domain of mathematics: trigonometry, backward digit span, and logic. The study employs a dynamic multimodal approach (i.e., physiology: electroencephalography or EEG; skin conductance or SC, and blood volume pulse or BVP; behavior: patterns of the learner interaction with the system; performance during a cognitive task; and two video cameras to record the learners' faces and the onscreen activity) to predict trends in the interaction experience, such as being stuck, off task, or in a state of flow. As predictive features, the model used the learner's individual differences and environmental factors (the current context and learner profile) as well as dynamic features that track the interaction experience over time (the temporal evolution of the learner's experience). Authors used a hierarchical probabilistic framework using a dynamic Bayesian network (DBN) to concomitantly predict the probability of each trend and the emotional responses that followed it. Participants completed the Big Five Inventory (BFI) questionnaire to assess learners' personality traits, whose answers served as the ground truth. The model was trained using 1848 samples (42 observations for each participant). A 10-fold cross-validation technique was employed to evaluate the model performance. The findings show that the proposed model outperforms static modeling approaches (e.g., static Bayesian networks or SBNs) and three non-hierarchical static algorithms (e.g., Naive Bayes classifiers, decision trees, and support vector machines). The DBN model achieved an accuracy of 82% to characterize a positive vs. a negative experience, as well as an accuracy ranging between 81% to 90% to predict four emotions related to interaction: stress, confusion, frustration, and boredom.

One limitation of this study is that the link between affect and learning outcomes was not explored.

Kai et al. (2015) conducted an online quasi-experimental (qualitative pre-/post-test testing physics knowledge and skills) classroom study to compare the performance (i.e., detection accuracy of the affective states: boredom, confusion, delight, engaged concentration, and frustration) of two types of affect detectors in a physics game, Physics Playground, which was designed to teach principles of Newtonian physics. In addition to affect, student behaviors were also coded as: on task (i.e., when looking at their own computer), on-task conversation (i.e., when conversing with other students about the task), and off-task (i.e., in other situations, such as using a cell phone). This study uses multimodal data sources. Six video-based detectors (i.e., facial expression recognition was applied to video data collected via a webcam) were compared with six interaction-based detectors (i.e., learners' interactions with the game were recorded in log files). They sampled n = 137 (80 females) Grade 8-9 US public-school students. Students were tested in groups of 20. They played Physics Playground during regular 55-minute class periods over four days. Live field observations of learners' on-task versus off-task behaviors were collected during gameplay by two Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP 2.0; Baker & Ocumpaugh, 2014) certified observers. Usually, BROMP is implemented using the Android app, Human Affect Recording Tool (HART; Baker et al., 2012). The observations constituted the ground truth affect and behavior annotations used in training automated detectors for both types of detectors to predict learners' affective states and off-task behaviors. There were fewer than 2087 observations of affect and behavior used in the creation of the detectors.

For the interaction-based detector, 76 gameplay features were extracted from the student interactions with the game (e.g., the time between the start and the end of a level, the mean number of gold and silver trophies obtained in a level, etc.). Several classification algorithms were built to detect each affective state using a two-class approach, where the affective state (e.g., engaged concentration) was discriminated from the combination of the rest of the instances (e.g., frustrated, bored, delighted, and confused combined, referred to as "all other"): JRip, J48 decision trees, KStar, Naïve-Bayes, step regression, and logistic regression. Behaviors were also grouped into two classes: off-task behaviors versus the not off-task behaviors (i.e., combined *on-task behaviors* and *on-task conversation behaviors* related to the game).

Model evaluation was conducted using a 10-fold batch cross-validation process on student data and the performance metric used was A' (i.e., the probability of correctly determining if an instance belongs to a certain affective state) or the area under the receiver operating characteristic (ROC) curve. The results revealed that classification was successful, exceeding chance for both data modalities. Moreover, video-based detectors (average A' of .695) slightly outperformed interaction-based detectors (average A' of .634), with video-based detectors showing a stronger advantage for *delight* (Kai et al., 2015). Although accuracy of the two detector suites was more comparable for the other constructs, the video-based detectors showed some advantages for *engaged concentration* and *frustration*, exceeding interaction-based detectors on 5 of the 6 constructs.

One limitation of this study is that the link between affect and learning was not explored.

Korn & Rees (2019) conducted an experimental study to evaluate the learning effects of performing repetitive tasks in gamified environments by collecting and analyzing bio-signals and performance measures.

They collected data from n = 23 (10 females) participants in Germany, of which n = 19 were students from the Offenburg University and n = 4 were trainees at the MAHLE International GmbH, with an average age of 23.6 years. Participants were randomly assigned into two groups: 13 participants who used the gamified application and 10 participants who used the applications without gamification. Participants were instructed to assemble 10 Lego houses using 12 bricks in 20 minutes. Two cameras were set up to record behavior, one in front of the student to record facial expressions and one over the participant's left shoulder to record the assembly area. The participants were also adjusted to a Shimmer wristband sensor on their right hand to track their electrodermal activity (EDA). To minimize noise in the EDA data, participants were instructed to use their left hand during the intervention.

Results showed that the gamification group spent less time completing the activity compared to the non-gamification group. However, the gamification group made 0.7 more mistakes on average when building the Lego houses than the non-gamification group. Results from the bio-signals showed that instances of joy were detected more for the gamification group than for the non-gamification group, supporting the authors' hypothesis that gamification increases positive emotions. The EDA analyses revealed that the gamification group tended to experience a consistent state of arousal, while the non-gamification group tended to experience boredom. Anger emotions were also identified more often in the gamification group, indicating that gamified environments increase emotions.

Lee-Cultura et al. (2020) conducted a quasi-experimental (within-subjects study) to identify how Avatar Self-Representations (ASR) influence the learners' affective state while playing Motion-Based Touchless Games (MBTG), because this style of games has become widely popular in multiple educational domains, such as literacy, STEM, social skills, and development of motor skills. The authors explain that literature that explores the relationship between learners' affective state, behavior while playing an educational game, and how they feel represented by a game's character is scarce and usually conducted qualitatively through interviews and expert observation. The researchers used multimodal data collected from different physiological channels, such as eye-tracking glasses, webcams, and wearable devices.

The study collected data from n = 46 (28 females) children with an average of 10.3 years of age. The study was conducted in two settings: a science center, where n = 30 children participated, and an elementary school, where n = 16 children participated. Each intervention consisted of 9 gameplay sessions that lasted on average 30 minutes. The sensors used included a pair of Tobii eye-tracking glasses and an Empatica E4 wristband to collect heart-rate variability (HRV), electrodermal activity (EDA), skin temperature, and blood volume pressure (BVP), while the videogame setting included a webcam, to collect videos of facial expressions, and a Kinect Skeleton, to collect movement data from 20 joints (e.g., head, shoulders, spine, hips, hands, knees, and feet). The learners played three times (i.e., practice round and two non-practice rounds) for each of the three levels of the ASR. The levels of the ASR can be described as follows: (1) low ASR, where the avatar was in the shape of a white-hand moved minimally, (2) moderate ASR, where the avatar was in the shape of a blue yeti and was controlled by the learner's movement, and (3) high ASR, where a photo-realistic avatar of the learner mimicked the student's complete range of movement.

Results from an ANOVA analysis, using the level of the ASR as independent value, demonstrated that arousal (measured by EDA) and stress (measured by skin temperature) were significantly different across the degrees of ASR. Arousal significantly increased as the ASR level increased, and stress significantly decreased as ASR level decreased. These results indicate that children were most engaged using high ASR and least engaged with low ASR because of the playfulness of the avatar and the intrigue of seeing themselves in the learning environment.

One limitation of the study was the age of the study population, given that slightly younger or older populations might produce different results and their needs in ASR might differ. Another limitation stemmed from the collection of physiological data because measuring children's experience via multimodal data involves a degree of error due to potential interference. Considerations of additional data-streams (audio and interviews) may have offered additional insights. Future work, including longitudinal studies, is needed to determine whether the findings hold true over time.

Leong (2015) conducted a quasi-experimental study to evaluate a logistic regression with lasso regularization model to student detect frustration while performing programming exercises using contextual (i.e., status of completion, number of submissions) features and keystrokes from a Java tutoring system developed by the author of the study. This study highlights the importance of detecting when a student might become frustrated while learning complex subjects, such as programming, and ultimately offer the student intervention before they lose motivation and abandon the task. The author explores the hypothesis that frustration can be detected with only contextual data and keystrokes, without the necessity of using expensive and intrusive physiological sensors.

This study sampled n = 14 (7 females) students from a post-secondary vocational institution in Singapore in a laboratory setting. All the students were enrolled in a diploma course in information technology and had at least one year of programming experience. The students were presented with instructional pages on six different topics (i.e., use of variables, loops, conditionals, and arrays) on the basics of the Java programming language. Each of them included two programming exercises, totaling in 12 exercises overall. The exercises consisted of code snippets with blank spaces for the students to complete the missing lines. The students were able to run the code with the help of a compiler that allowed them to check their answers and fix mistakes, if required. When the students were satisfied with their answer, they could submit the final answer and the platform would mark it as complete. The interventions were on average 81.3 minutes long.

Students' keystrokes, contextual features, and timestamps were recorded and used to build the logistic regression model. The outcome of this study was a binary output (i.e., whether the student was frustrated or not). The dataset was limited (n = 56), making the model prone to overfitting, which was addressed by employing a regularization technique to penalize large coefficient values (Peng et al., 2002). Two models were compared: (1) a model that only considered the contextual behavior, yielding an accuracy of 58.3% and (2) a model that combined the contextual behavior and the keystrokes, yielding an accuracy of 66.7%. Although the multimodal model yielded a low accuracy, the recall measure was high (83.3%), meaning that the model was highly accurate when making the correct prediction.

In sum, the results support the hypothesis that frustration of a student with the programming task on hand can be detected through utilizing the contextual and keystroke features of the student collected within the tutoring system.

One of the limitations of this study is that the approach used is not quite useful to detect frustration in real-time but can help the instructor identify which of the topics or exercises are frustrating the students. The author suggests that implementing new channels of data (i.e., clickstreams), could be useful to create a more robust platform that offers instructional aids (i.e., hints) to struggling students.

Mangaroska et al. (2020) conducted a quasi-experimental study to determine how using multimodal data, in addition to using IDE-log data, could improve machine learning models' interpretation of learner behaviors.

The study collected data from n = 46 (8 females) university students enrolled in CS majors at the Norwegian University of Science and Technology. All the students had programming experience in Java using the Eclipse IDE. The authors collected for channels of data: gaze data using a Tobii X3-120 eye-tracking device mounted at the bottom of a computer screen, physiological data (i.e., HR, BVP, temperature, and EDA using the Empatica E4 wristband), facial expressions using a LogiTech web camera pointing directly to the participants' faces, and log data from the IDE using an Eclipse plugin that captured the moments when students saved their programs. The intervention started with setting up and calibrating the sensors per participant. During the intervention, the students were asked to finish three small debugging tasks with different difficulty levels (i.e., easy, medium, high) within 20 minutes. This task was used as a level placement activity. Afterwards, participants had 40 minutes to solve the main task, where they debugged code from a Person class method in Java, and five debugging sub-tasks as questions, where they had to fix the code to make the class successfully perform parent-child relationships.

Eight Random Forest classifier models were trained to observe how physiological data improves the base model trained with only IDE log data. The model was evaluated using 10-fold cross-validation. For each of the models, accuracy, precision, recall, and F1-Score were calculated. To compare the models, the adjusted R^2 metric was considered. Results from the adjusted R^2 showed that the second model (i.e., log data and eye-tracking data) and the eighth model (i.e., log data, eye-tracking data, facial expressions, and physiological data) significantly outperformed the base model. Feature performance analysis was conducted. From the log data, expertise was the most influential variable, indicating that log data can imply patterns that represent previous knowledge on how to perform the debugging tasks. From the eye-tracking data, the average length of the saccade was the most influential variable, indicating an increase in mental effort. For the facial data, nose wrinkling was the most influential data. Finally, for the physiological data, average temperature was the most influential variable.

One of the limitations of this study is that they do not exclude the possibility of measuring complex internal conditions with other methods such as think-aloud protocols, pre-/post-tests, or self-report questionnaires. Additionally, findings originate from tasks based on problem-solving practiced by an individual using a computer screen, which can be difficult to generalize. A limitation on the study design is that it was performed in a controlled environment and students were aware of the study. Lastly, a limitation on the data analysis was that the authors only used one algorithm; using different algorithms may produce different results.

Mills et al. (2017) conducted a quasi-experimental (within-subjects) study to evaluate a model to detect cognitive load on learners using and electroencephalography (EEG) system based on the cognitive load theory, that suggests that learning can be promoted or prevented by the amount of cognitive load produced by the system (Paas & Ayres, 2014). The EEG measures the brain's voltage that passes through the scalp, which may be an index of cognitive state.

The study collected data from n = 12 (7 females) high-school students enrolled in a Grade 9 biology class. EEG data was collected using the QUASAR, a hat-like headset that uses ultra-high impedance dry electrodes. Guru is a dialogue-based ITS where a digital tutor interacts with the student to instruct them on 120 biology topics in 15 to 40-minute sessions. Two levels of difficulty were developed to fit the research design, where the versions of the topics, Ground

Building Instruction (CGB) and the Scaffolded Dialogue, were mixed up. The intervention started with setting up the headset on the student and ensuring that the electrodes were properly producing data. The students then completed the training tasks of the ITS. Afterwards, they completed a knowledge pre-test before completing a session of the biology topic. This process was repeated twice. After each session, the students completed a survey to assess the difficulty of the material and a knowledge post-test. After the second session, the students completed an eyes-closed and eyes-open tasks. The intervention lasted on average 1.5 hours per student.

Four models to detect cognitive states were trained using the Qstates algorithm. The output was determined using a stratified k-fold (k = 6) cross-validation technique. Multivariate normal probability density functions (MVNPDF) were used to classify the instance as low or high cognitive load. Linear mixed effects regression analyses were performed to evaluate the models. Results showed that students did not perceive the difficulty changes. Further analyses on the Scaffolding phase were performed and found that cognitive load was higher during the difficult scaffolding questions. These results can be corroborated by the students' rating from the post-intervention survey.

One of the limitations from this study was that the sample size was small (n = 12). A second limitation was the setting of the study. Being a laboratory environment can make it difficult to generalize to a classroom setting.

Muñoz et al. (2016) conducted a quasi-experimental study (pre-/post-test testing physics knowledge and skills) sampling n = 118 undergraduate students enrolled in an engineering degree at the ITESM in Mexico. The study evaluated a Bayesian Networks model that detected the emotional state of students while interacting with a serious game, PlayPhysics. The model is based on Pekrun et al.'s (2007) control-value theory of achievement of emotions, which posits that academic emotions are influential to learners' motivation to learn, therefore affecting their academic outcomes. However, the authors emphasize how most of the modeling of affect in education is based on theories from other fields, such as neuroscience. Therefore, pointing out the novelty of their study where they attempt to detect the students' emotions using a game-based environment that can emotionally connect with the learner via storytelling.

After completing a pre-test, the students interacted with the first challenge of PlayPhysics, this challenge was about the one-dimensional multilinear motion. During this challenge, the students were required to choose the correct direction of the in-game vehicle, correctly set the mass of the object, consider the amount of fuel of the vehicle, and define the appropriate braking time. Finally, they completed a post-test and a qualitative questionnaire. Students self-reported their emotional state before, during, and after performing the game activity. During the interaction with the game challenge, the student's emotion could be reported at any time, using the EmoReport wheel. At the end of each game dialogue, students self-reported their emotions. The researchers collected multimodal data from the game: contextual behavior of the students with the game (i.e., mouse position, number of times they asked for help) and behavioral data from the self-report surveys of the learners' emotion.

The model was created using a dynamic Bayesian Network (DBN) to model a temporal relationship (Neapolitan & Jiang, 2007). In this case, there were three networks: (1) prospective outcome, corresponding to the emotions observed related to the future activity; (2) activity, corresponding to the emotions observed when the student was interacting with the game; and (3) retrospective outcome, corresponding to the willingness to keep interacting with the game. WEKA was employed to perform random sampling and to convert continuous variables into categorical variables. Model evaluation was conducted using a 10-fold cross-validation process on student data. The prospective-outcome emotions (i.e., anxiety and hope) were classified with 80% and 67% accuracy respectively. The activity emotions (i.e., enjoyment and frustration) were classified with 67.8% and 60%, respectively. Finally, the retrospective-outcome emotions (i.e., anger and gratitude) were classified with 77% and 0%, respectively.

In sum, results showed that the model attains fair-moderate accuracy with results that are not random using answers in game dialogues and contextual variables, but it is not highly accurate. The authors suggest that future work involving observable features, such as video recordings of facial expressions and audio could help improve the classification accuracy of the model.

Park et al. (2018) conducted an experimental (pre-/post-test testing programming knowledge and skills) study to analyze how learners engage with embedded Python programming exercises using Elicast, a screencast tool that allows instructors to embed programming activities within the video recordings of the lecture. On the student side, the video recording is displayed and when the exercise is presented, the video pauses, allowing the student to complete or skip the activity. Elicast provides immediate feedback to the student and the system can verify that the functionality of the students' submission, even if incomplete, is the same as the correct solution.

This study sampled n = 63 undergraduate students with previous introductory knowledge of Python. The majority of the students (n = 46/63) had only taken one course on CS before the experiment. The experiment consisted of three stages. First, students completed a pre-test survey asking the students their proficiency in programming, this survey was based on Wigfield & Eccles's (2000) Expectancy-Value Theory of Achievement Motivation. Afterwards. Second, the students watched two video lectures on Elicast, one lecture that included an embedded exercise and a second lecture without an embedded exercise. The purpose of this was to analyze the effectiveness of embedded exercises. Before each lecture, the students were asked to answer the five-question programming skills pre-test about the concepts to be covered during the lecture (i.e., Introduction to Python, Queues, and Regular Expressions). Finally, the students answered a skills post-test and an open-ended survey to report their satisfaction and how effective they considered the embedded exercises. Each lecture took around 15 to 20 minutes, making the overall process about one hour long. To analyze the platform, the authors applied a combination of qualitative data from the post-test and observations, quantitative data from the pre- and post-tests, and the platform's logs from the lectures that included embedded exercises to analyze active engagement. There were 2612 video navigation events of video engagement. Unequal variances t-tests were used to measure the learning gains before and after the procedure.

Results show that students actively engaged in lectures when the lectures have embedded programming exercises. The answers to the questions from the post-study survey showed consistent results of active engagement of students. In the post-study survey, most of the students reported that the embedded exercises positively affected their learning experience. From the free form questions, 13 students mentioned that they were able to stay focused and be engaged throughout the lecture because of the embedded exercises. On the other hand, a few students felt disengaged from the lecture because there were too many things to do. One of the limitations of this study was that the lectures selected for this intervention were already familiar to the majority of participants, which made it difficult to determine the effect size of learning gain with the pre-test post-test experiment.

Pham & Wang (2016) conducted an experimental study that sampled n = 32 (9 females) University students. It used *AttentiveReview*, an affect-aware, adaptive intelligent tutoring system (ITS) designed as an intervention technology easily integrated into mobile massive open online courses (MOOCs) to enhance learning through adaptive content review. The system employed one data modality, being only based on physiological signals. *AttentiveReview* uses the built-in camera of unmodified smartphones and on-lens finger gestures for video control while participants watch MOOC tutorial videos on their smartphones. It blends the photoplethysmography (PPG) sensing process with the video via a tangible video control channel. Specifically, participants cover the back camera lens of a smartphone to play a tutorial video and uncover the camera lens to pause the video. The transparency detected in the learner fingertip by the camera changes with every heartbeat as the heart pumps blood to capillary vessels and it correlates with heart beats. Thus, the implicit PPG signals have been used in the literature to infer learners' affective states (Pham & Wang, 2016). The goal of *AttentiveReview* is twofold. First, it aims to predict a learner's perceived difficulty levels of learning each topic using input features extracted from rich but noisy physiological signals such as PPG sensing (i.e., waveforms captured implicitly from fingertip transparency changes) on unmodified smartphones via a back camera. Second, it adaptively recommends the optimal review materials through personalized review sessions based on a user-independent model.

The findings show that *AttentiveReview* was able to capture PPG signals effectively to provide adaptive review (i.e., recommend review materials) that improved learners' information recall (+14.6%) and learning outcomes (+17.4%) compared with the no review condition. Also, *AttentiveReview* achieved comparable performances with significantly less time when compared with the full review condition. Participants reported that *AttentiveReview* was intuitive and responsive. The authors used a supervised learning technique, linear kernel ranking support vector machine (SVM) to predict perceived difficulty in each learning topic from PPG signals. Specifically, *AttentiveReview* extracted 17 temporal domain features and frequency domain features from a learner's PPG waveforms that were collected from the learning process. The model yielded an accuracy of 62.5% in predicting perceived difficulty. Participants completed a survey to indicate their perceived difficulty of each topic in the video tutorials. This information provided the ground truth labels for the *AttentiveReview* model.

One of the limitations of this study is that the system requires course-dependent training. Additionally, the review recommendations are at the learning topic level and some MOOCs may not organize their tutorial videos by learning topic. This may limit the prediction accuracy of the model especially for short videos. Also, the recommendation of review materials takes place after finishing multiple learning topics (due to the ranking SVM algorithm chosen in this study) instead of during the learning process. Lastly, the link between affect and learning was not explored.

Psaltis et al. (2017) conducted a quasi-experimental study to evaluate neural network models using multimodal data to detect student engagement while playing the self-developed prosocial game, Path of Trust. In the game, the learner chooses one of the playable characters: the Muscle who follows directions and the Guide who interacts only through suggestions. The two characters must work together to collect equal parts of the treasure. A Microsoft Kinect device was used to collect the affective data: facial expressions were collected using the Kinect's SDK face tracking engine and the Kinect sensor was used to extract joint-oriented skeleton tracking.

The study collected data from n = 72 (34 female) primary school students from three different schools in Greece. The intervention consisted of two sessions where the students played the Path of Trust game in a classroom setting. The setup consisted of one desktop PC with a Microsoft Kinect sensor. After each of the gameplay sessions, the students answered the 19-

question Game Engagement Questionnaire (GEQ), to quantify the learners' engagement when playing a game (Brockmyer et al., 2009). The intervention time was between ten and fifteen minutes.

To detect affective states, a two-layered network model with seven stacked Artificial Neural Networks (ANNs), six at the first layer and one in the second layer, was trained. This model was compared with other multimodal affective states classifiers (i.e., Linear Weighted, Nonlinear SVM, and Shallow NN). The emotions recorded in the dataset were labeled as anger, fear, happiness, sadness, surprise, and neutral (i.e., where none of the previous emotions were detected). The dataset contained n = 750 three-second videos from n = 15 subjects. The model was trained using an augmented noisy dataset with added neutral state samples. The proposed neural network outperformed the other classifiers with a recognition rate of 98.3% when combining the two data channels.

To detect engagement, a binary classifier neural network was trained. The model obtained an average engagement value of 0.728, meaning that a game that is challenging can trigger high levels of engagement to the players. Compared to the other classifiers, the neural network outperformed them with an 85% classification rate.

Rajendran et al. (2019) conducted an experimental study to evaluate a model that detects students' frustration in real-time using ITS log files to offer immediate personalized solutions, such as providing motivational messages as feedback. The model was implemented in the software Ei MindSpark (Rajendran & Muralidharan, 2013), an AI powered platform for learning mathematics. The researchers added motivational messages based on the identified cases that caused frustration.

The study collected data from n = 769 Grade 6 students from three schools (School 1: n = 326; School 2: n = 279; School 3: n = 164) from different cities in India (Rajkot, Bangalore, and Lucknow). After preprocessing and data cleaning, the dataset was reduced to n = 188 students in the experimental condition (i.e., receiving motivational messages from the ITS) and n = 188 in the control condition (i.e., not receiving motivational messages from the ITS).

To evaluate the model, the researchers conducted a Mann-Whitney test to compare the number of instances of detected frustration between the control group, which were collected a week prior from the intervention, and the experimental group. Results from this test indicated that the average number of instances of frustration was greater from the control group than the experimental group. In other words, when the students receive immediate feedback, in this case through motivational messages, the students appear to become less frustrated when using the Ei MindSpark platform.

In sum, the results from the statistical analysis show that providing students with motivational messages, which addressed the cause of frustration when using an ITS, significantly reduced frustration in mathematics learners.

Sharma et al. (2018) conducted an experimental study to understand the relationship between children learning basic programming concepts and the behaviors and attitudes presented while coding using eye gaze data. The authors highlight the relevance of interpreting the learners' affective state while learning to code because of the demand for younger students to learn this topic that is usually deemed as difficult.

The study collected data from n = 44 (12 females) Grade 8 to Grade 12 students with an average age of 12.64 years, recruited from local schools. The intervention consisted of five coding workshops at the Norwegian University of Science and Technology in Trondheim, Norway. Each workshop lasted approximately four hours. For the first part of the intervention, the students interacted with digital robots, and they received a paper tutorial to learn how robots react to simple loops using Scratch for Arduino. Next, the students developed a simple game using Scratch. Researchers collected eye gaze data during both workshop activities using four SMI and one Tobii eye-tracking glasses. After the intervention, a 5-point Likert-scale survey asked the learners to rate their perceived learning, enjoyment, teamwork, and intention to participate.

Results from correlations analyses demonstrate that gaze behavior moderate the relationship between the explored attitudes. In the relationship between teamwork and perceived learning, participants working on teams with high gaze similarity reported higher levels of perceived learning. In the relationship between teamwork and enjoyment, participants with high gaze similarity reported higher levels of enjoyment. In the relationship between intention to learn and perceived learning, participants with gaze similarity reported higher levels of perceived learning.

Sharma et al. (2021) conducted an experimental study to investigate the relationship between the joint gaze of students coding in pairs and their affective state while they were performing block-based programming tasks. The authors present a system that uses eye-gaze behavior and facial expressions collected from a Logitech webcam, and logs from Scratch to assess the learners' emotion and performance when working in teams.

The study collected data from n = 50 (29 females) Grade 8 to Grade 10 students who voluntarily attended an after-school workshop at the Norwegian University of Science and Technology. The students were organized into 10 pairs and 10 triads. The learners were introduced to block-based programming through Scratch and were instructed to modify and develop their own games by iterative coding and testing and by working either in pairs or in triads. After completing the games, the teams reflected on the process and played each other's games. Several measurements were collected to evaluate the system. The Joint Visual Attention (JVA) measurement represents the time the team individuals spent looking at similar objects within a timeframe computed from facial video. The Joint Emotional State (JES) measurement represents the time the team individuals spent experiencing the same emotion (i.e., frustration, boredom, or confusion) computed from facial video. Finally, the Information Flow measurement represents the amount of information presented to the students on the screen, computed from the screen recordings.

Results from correlational analyses showed no significant differences between the number of members in a team (i.e., the number of students in a team did not affect performance). All of the results are based on a collaborative coding activity context. Groups of children that performed well spent time looking at similar locations on their screen, indicating strong cognitive load. On the other hand, the teams that had lower agreement when looking at similar locations on their screen displayed lower performance. When analyzing instances of boredom between high and low performing students, the results indicate that students that showed joint states of boredom displayed lower performance. Similar results were found when analyzing states of confusion, meaning that students who showed joint confusion also displayed lower performance. Groups of children that performed poorly experienced confusion together. When analyzing instances of confusion, the arrangement of the information presented on the screen influences higher joint states of confusion. In other words, when the team is confused together, it is unlikely that the team members will produce high quality code, which can lead to boredom and/or frustration in later stages. Finally, the low performing group showed higher instances of frustration. The overall result of this study implies that JVA causes cognitive load for high gain teams. This result can indicate that students arranged in teams can produce more correct code when in JVA (i.e., the time the students spent looking at similar locations on the screen) because it is correlated with mutual understanding and high collaboration quality.

One limitation of the study is that findings might be restricted by the fact that the tasks were tailored to the study. Also, participants were sampled from schools where students already showed an interest in the workshops, so self-selection may be a concern. Finally, all of the data collected came from only one session, thus complicating the generalization of the results.

Sinha et al. (2015) conducted an experimental study to evaluate a model that predicts learners' cognitive flow using multimodal physiological data. Cognitive flow is defined as the mental state where the learner is fully concentrated while feeling involved and enjoying the activity. The researchers developed a modified version of the Tetris game where the falling objects are replaced with Stroop color-texts that the participants needed to assign to color boxes at the bottom of the screen. The authors highlight that creating learning platforms where the learners find themselves in a constant state of flow is challenging, however doing so can ultimately improve learning experiences and outcomes.

The study collected data from n = 20 (10 females) right-handed engineers from the researchers' lab with an average age of 30 years. The researchers collected the following channels of data: electroencephalogram (EEG) signals using the NeuroSky device placed on their left earlobe, electrodermal activity (EDA) using the eSense device placed on their middle and ring left hand fingers, oxygen saturation and pulse rate using a Contec oximeter placed on their left index finger, and keystroke data. Tasks in the game were designed to induce flow or boredom. Participants were randomly assigned into one of the two categories and then alternated to the opposite one: the order of the tasks was boredom-flow-boredom-flow or flow-boredom-flow-boredom. The intervention consisted of four blocks. For two of the blocks, the participants interacted with the game in one of the categories (i.e., boredom-flow-boredom-flow), whereas for the last two blocks, the participants interacted with the game in the opposite category (i.e., flow-boredom-flow-boredom). After each block, the participants answered the Game Flow Inventory (GFI) questionnaire to measure their affective state (i.e., level of engagement, enjoyment, or happiness) and motivation.

The results of this study were divided by data channel. Results from the subject feedback using the GFI survey were evaluated using a t-test, which showed that 16 of the 20 subjects experienced a flow state. Results from the keystroke analysis were standardized. For the boredom condition, most values of the correct number of keystrokes divided by the total number of keystrokes were close to 1, whereas for the flow condition, the value of correctness decreases as the speed increases. Results from the EEG analysis were estimated by a Gaussian Mixture Model (GMM) and showed no significant differentiation for the participants. Results from EDA data were evaluated by a t-test and did not show significant differences between conditions.

Sottilare & Proctor (2012) conducted a quasi-experimental study to find methods to allow an Intelligent Tutoring System (ITS) to detect a leaner affective state using student behavior and physiological responses. The rationale of this study is for ITS to understand the affective state of the learner and adapt the instructional strategies to improve learning experiences. The target population of this study was military students who interacted with the training package, Tactical Combat Casualty Care (TC3), a software used to train cadets on hemorrhage control.

The study collected data from n =124 (16 females) cadets from the United States Military Academy, with low-moderate competence in tactical combat casualty care. The intervention started with a demographics survey, a training course on the software, a knowledge pre-test, a mood assessment before the intervention using the Self-Assessment Manikin survey to measure pleasure, arousal, and dominance related to the learners' affective state (Bradley & Lang, 1994). During the intervention, the learners interacted with the TC3 to apply the knowledge learned during the training course. After the intervention, the students answered a knowledge post-test, Self-Assessment Manikin. The researchers collected timestamped logs from the platform, and interactive controls to detect strategy.

Linear regression analyses were conducted to find relationships between the variables that predicted the following moods: (1) pleasure, (2) arousal, and (3) dominance. Results of the pleasure and dominance detecting models showed that there were no reliable predictors. These results were unexpected to the authors because it was hypothesized that mouse movement would be a reliable arousal predictor. To measure the differences between the mood of the learner before and after the intervention, a non-directional t-test was conducted. For the pleasure and arousal moods, significant differences were found. However, there were no significant differences in dominance for the dominance mood. Finally, to measure performance, a linear regression analysis was conducted. Results showed that initial dominance, mouse movement, and final knowledge scores explained a large portion of the variance on performance ($R^2 = 0.23$). These results were not expected for the authors because it was hypothesized that previous training experience and interest in the topic would be reliable performance predictors.

Srivastava et al. (2018) conducted a quasi-experimental study to compare the performance of machine learning classifiers using mid-level gaze eye activity to detect the type of task the learner is performing on a computer screen, using a Tobii Pro X2-30 eye tracker. Mid-level gazes are defined as a combination of low and high-level gaze features that are shape- and distance-based and that do not require information of the interface design.

The study collected data from n = 24 (8 females) postgraduate students and research staff from a university, with an average age of 29.8 years. Participants were proficient C# or Python programmers. Before the intervention, the participants were fitted with an eye-tracker and a posture-sensing chair in a laboratory environment. During the intervention, each participant performed five common desktop activities (read, watch, browse, play, and search) and three software engineering activities (interpret, debug, and write). Depending on the activity, the participants were instructed to interact with the keyboard and/or mouse. Each session lasted 60 minutes on average.

Using only low-level gaze or a combination of low and mid-level gazes, three classifier models were trained using the following algorithms: SVM, K-NN, and Random Forest. The models were evaluated using 4-fold cross validation. Afterwards, hyperparameter tuning was conducted to find the best hyperparameter values for the models. The SVM classifier C parameter (to control error) was set to 10 and the gamma parameter (to give curvature weight of the decision boundary) was set to 0.01 with a Radial Basis Function (RBF) kernel. The Random Forest classifier was tuned using 1000 trees. Finally, the K-NN model used k = 10 neighbors.

The results showed that the Random Forest using the combination of gaze features classifier outperformed the rest of the models, yielding the highest F1-Score. These results suggest that mid-level gaze features are related to the type of activity.

Standen et al. (2020) conducted a quasi-experimental study evaluate the online platform MaTHiSiS, an adaptive learning system used to identify students with intellectual disabilities

(ID) and affective states linked to learning (e.g., engagement, frustration, and boredom) using multimodal data. The authors highlight the importance of developing personalized and accessible learning systems to address the limitations of current standardized platforms for learners with ID.

The study collected data from n = 67 (21 females) students, ages 6 to 18 years, with an ID or with an autistic spectrum condition (ASC) from six schools in the UK, Italy, and Spain. Three participant groups were identified: (1) those with ID only (n = 23); (2) those with ID and autistic tendencies (n = 22); and (3) those with the primary diagnosis of autism (n = 22). The intervention used an online equivalent of a specific lesson in traditional learning environments, where several learning goals are defined and are expected to be acquired. Interaction with the system was made by devices (laptop, tablet, or NAO robot) at the instructor's discretion.

Results from the model indicate that sensor data can identify the three different affective states (engaged, frustrated, and bored) all with a strong relationship with achievement. "Lack of boredom" was the state most strongly linked to achievement. There was significantly more engagement and less boredom in the intervention than in the control sessions, but no significant difference in achievement.

Subburaj et al. (2020) conducted a quasi-experimental study to evaluate the performance of models using different channels of multimodal data that predict student success in solving a level in Physics Playground, an educational game to learn Newtonian physics concepts. The novelty of this project is that it adds a layer to the multimodal studies by exploring multiparty signals (i.e., combining signals of individuals arranged in teams). However, this approach raised questions on how individual features could be weighted when building machine learning models. The researchers focused on collecting nonverbal behavior, such as facial expressions, acoustic-prosodic information, eye gaze, and task information.

The study collected data from n = 303 (56% female) undergraduate students from two large public universities (38.5% from one university) in the United States. Students were arranged into 101 groups of three. The students met using the Zoom video conferencing software. Before the intervention using the educational game, the students answered a demographics survey, and the following assessments: (1) a validated measure of physics self-efficacy, to assess their ability in physics; (2) the Big Five inventory, to assess personality dimensions (i.e., extraversion, agreeableness, openness, conscientiousness, neuroticism; Gosling et al., 2003); (3) the Leadership Domain Identification measure, to assess their self-perceived capability of leadership; and (4) the Individual Satisfaction with the Team Scale, to assess willingness to work in teams, and teamwork self-efficacy (i.e., their self-perceived ability to work in teams). Finally, they completed a tutorial on the platform before the intervention. For the intervention, the students met via Zoom using a personal computer with a webcam to record facial expressions and upper body posture, an audio recorder for dialogue, and a Tobii 4C eye gaze recorder. The teams interacted with the game for three 15-minute blocks. For each block, one student was randomly assigned as the leader or controller who was in charge of the mouse interactions, whereas the two other students were observers who could contribute to the solution for that level.

The authors trained Random Forest models to predict successful or unsuccessful attempts to pass the level in the game. The model was evaluated using five-fold nested cross validation, whereas the training set was split into three folds for hyperparameter tuning using grid search to tune the number of trees in the forest and the maximum depth of the trees. The performance metric to assess the model was AUC-ROC. Models were trained using different sets of their data (i.e., unimodal, and multimodal using different combinations). Based on the performance results, the best model (AUC-ROC = 0.73) was the one that combined eye-gaze behavior, task context, and facial expressions, excluding the acoustic prosodic features. After this result, models to compare the performances of nonverbal and language-based models were trained using Random Forests. Results showed that the nonverbal model was the best model (AUC-ROC = 0.73).

The & Mavrikis (2016) conducted an experimental study to determine how different patterns of eye fixations and saccades of students learning programming in Codecademy relate to their achievement and performance. The authors highlight how the use of online learning systems has improved learning experiences by helping educators monitor progress and deepening learning for students by facilitating problem-solving activities. Thus, addressing the different learning approaches using bio-signals, in this case eye behavior, can help educators understand the effectiveness of personalized online learning systems.

The study collected data from n = 60 final year students from the School of Information Technology at the Nanyang Polytechnic in Singapore. Before the intervention, the eye-tracking device from Eye Tribe was calibrated. The intervention consisted of completing 13 programming tasks in the PHP language provided by Codecademy. Video-screen recordings were documented to validate the eye-tracking recordings.

Analysis of the eye-tracker data identified three types of visual scanning behaviors: (1) Type FP1, where the eye fixations were consistent among all the blocks (i.e., Introduction, Instructions, Hints, Editor, and Output); (2) Type FP2, where the eye fixations were mostly on the Hints section and barely any eye fixations on the Introduction and Instructions sections; and (3) Type FP3, where the eye fixations were consistent on all blocks except the Hints section. Results from linear regression analyses revealed a strong positive relationship between the identified visual scanning behaviors and levels of engagement. These results highlight the relevance of creating models that identify the students struggling to learn the material (e.g., before assessments at the end of the semester) to provide additional support to those who are struggling with the material. Another significant result suggests that students who were consistently engaged with the system scored higher in online and traditional methods of learning.

Vail et al. (2016) conducted a quasi-experimental study to evaluate a model using facial expressions to find how genders differ while learning. Facial expressions have been widely analyzed to detect everyday emotions and with the growing literature on the relationship between emotions and learnings, this channel of data has been found to be informative when used to detect learning emotions (e.g., engagement, frustration, etc.). Literature has mentioned that females respond with more pronounced facial expressions, thus the novelty of this study is to find the extent of information females' facial expressions can provide.

The study collected data from n = 67 (24 females) undergraduate students enrolled on an introductory engineering course. The students interacted with JavaTutor, a tutorial interface to learn introductory computer science, and received texts from a human tutor when needed. The intervention consisted of six 40-minute lessons over four weeks. Before the intervention, students answered a content-based pre-test, and after the intervention, they answered an identical post-test. Facial expression data were collected using a Kinect depth camera, a webcam, and a skin conductance bracelet.

A standard *t*-test with the Bonferroni correction was performed to find facial expression differences between genders. Although no significant differences were found, females displayed a lower facial expression significantly more than males, while males displayed brow lowering and lip fidgeting more than females. Additional results found that certain facial expressions were associated with learning depending on gender. For example, females exhibited brow raising and nose wrinkling and males exhibited eyelid raising.

VanLehn et al. (2017) conducted an experimental study to evaluate the effect of a non-cognitive learning companion (LC) that reacted to students' affective states and provided motivational and affective messages to improve learning outcomes by inviting the student to persist. To detect affective states two sensors were used: a video camera to record facial expressions and a posture-sensing chair. Additionally, log data from the tutoring system were recorded.

The study collected data from n = 66 university students from Arizona State University, with an age range of 18 to 21 years. After filling out a background questionnaire and setting up the sensors, the students started the intervention which consisted of three parts. The first part was the introduction, where the participants studied 76 slides with simple exercises to learn how to use the system and learn introductory concepts of model construction. The second and third parts were the training phases, where the students solved as many problems as possible with the opportunity to refer to the slides from the previous phase; the students were reminded to keep talking. These two phases lasted 75 minutes. Finally, after a 30-minute break, the learning transfer phase started, where the student solved as many problems as possible without the support of the LC. The students were randomly assigned to one of three conditions: (1) the control condition, where the tutor only intervened during the training phase (e.g., offering feedback, hints, and instructions addressing the domain knowledge); (2) the meta-tutor condition, where the meta-tutor only intervened during the training phase (e.g., offering instruction on the learning strategy); and (3) the non-cognitive LC condition, where in addition to the tutoring and meta-tutoring, the LC intervened between tasks.

Results from the experiment showed that during the training phase, the meta-tutored students outperformed those who only had the tutor, but the advantage disappeared during the transfer phase. This result is consistent with findings of earlier studies. Results from the second experiment showed that the LC students outperformed the meta-tutored students during the training phase; however, during the transfer phase, there were no significant differences and there were no medium-sized differences in performance between the two conditions.
One of the limitations was the small sample for the second experiment which somewhat underpowered the performance results. Additionally, physiological measures had poor accuracy because they were calibrated using a different population.

Veliyath et al. (2019) conducted a quasi-experimental study to evaluate the performance of a model that detects students' attention in a classroom setting using eye gaze behavior. The researchers focused on gaze data as a method to estimate attention because it is almost non-intrusive to the subject and requires small amounts of setup, thus producing fewer errors in the data. Additionally, the authors highlight the relevance of detecting affective states, and attention in this case, to keep a student engaged, thus more open to learning and to improving learning outcomes.

The study collected data from n = 10 (4 females) junior undergraduate students enrolled in the field of mechanical engineering. The intervention took place in a computer-lab where the professor was located at the front of the class delivering a PowerPoint presentation. The students could use the computers, notebooks, and other tools freely. The computer in front of every student was set up with a Tobii 4C eye tracker and software installed to collect data of the student's interaction with the computer (e.g., timestamps, the application the student was using, etc.). Finally, a pop-up appeared every five minutes asking the students to rate the level of engagement of the previous five minutes of the class in a 10-point Likert-scale format. The purpose of the setup was to allow the students to decide to pay attention to the instructor or to follow along with the material on the computer.

Four binary classifiers to estimate attention were trained: Random Forest, Support Vector Machine, Adaptive Boosting, and Extreme Gradient Boosting. For each of the models, accuracy, precision, recall, and F1-Score metrics were calculated. The models were evaluated using K-fold cross validation (k = 20). Accuracy results showed that the Extreme Gradient Boosting algorithm outperformed the rest of the models (0.77). Feature importance analysis indicated that timestamp was the most significant feature (46%) of the model, followed by gaze location (31%), which can greatly influence the estimation of attention.

In sum, data collected from the eye tracker can be used to predict a student's attention as a measure of affect over the course of a class. However, one limitation of the study could be the existence of the Hawthorne Effect (Allen & Davis, 2011), whereby students might act differently when they are aware of the experiment.

Wu et al. (2020) conducted an experimental study to compare the statistical differences between students learning on SURGE, a physics digital game-based learning (DGBL) environment, and a traditional e-learning platform. Physiological data was collected to detect learners' attention level, using NeuroSky; affective state was collected using the emWave sensor; and cognitive load was measured using an eye-tracker. The authors make note of how game-based learning can increase learners' motivation and improve learning experiences and outcomes. However, the literature on the topic reported ambiguous findings regarding learning achievement.

The study collected data from n = 32 (18 females) university students in Taiwan. Before the intervention, the students completed the 30-question FCI pre-test, to assess the student's understanding of basic Newtonian physics concepts (e.g., kinematics, type of forces and the first three Newton's Laws). Students were randomly assigned to the two groups: the DBGL group or the traditional e-learning group, where both groups interacted with their respective platform for approximately ten minutes. After the intervention, the learners answered a post-test with two problems varying in difficulty from the Mechanics Baseline Test, a companion to the pre-test.

Results from a one-way ANOVA analysis showed that the DBGL group had a higher attention score than the static e-learning group, but no significant differences were found. Learners who start with a successful motivation exhibited a higher attention in game-based learning. No evidence on whether students pay more attention during game-based learning. The DBGL group did not have better affective experiences compared to the traditional group and had lower scores in total fixation durations and number of fixations, but higher scores on average fixation duration and percentage of viewing time (greater cognitive load). Finally, the DBGL group had better academic achievement in the post-test, although there was no statistically significant difference.

Xiao & Wang (2015) conducted a quasi-experimental study to report the efficacy of their proposed platform *AttentiveLearner*, a mobile learning system that captures learners' physiological states in Massive Open Online Courses (MOOCs) through heart rate sensing by having the learner cover the back-camera lens with their finger while playing a lecture video. This physiological data is a real-time feedback response on the learners' cognitive state.

The study collected data from n =18 (7 females) undergraduate and graduate students in a laboratory-based setting. The intervention consisted of an introduction to the system, afterwards, the students watch the introductory chapter of an MOOC on Game Theory, that included four chapters (i.e., Introduction to Game Theory and the Predator Prey Example, Normal Form Definitions, Dominance, and Nash Equilibrium). The total duration of the chapter was approximately fifty minutes. The students watched the videos on a Nexus 5 smartphone on landscape mode, with the freedom of pausing when needed. After each lecture, the students answered a survey about their interest on the topic and confusion levels in a 5-point Likert scale. Finally, the students completed a feedback questionnaire after each lecture.

From quantitative and qualitative data, the researchers build a dataset of n = 428 samples of interest and boredom predictions, as well as n = 490 instances of confusion. Later, they trained five supervised machine learning algorithms to predict the students' affective state. The algorithms evaluated were: kNN, Naïve Bayes, Decision Tree, Linear SVM, and radial basis function kernel SVM (RBF-kernel SVM) and used WEKA to train the classifiers. Leave-one-out cross-validation was used to evaluate the models. To determine the best algorithm, the average Kappa value was evaluated.

Performance results show that the RBF-kernel SVM yielded the best overall Kappa (0.297 and 0.269) when predicting boredom and confusion. However, one limitation of this approach is that the classifiers are binary and may not be sensitive enough at discerning other affective states, such as joy and engagement.

Xiao & Wang (2016) conducted an experimental study to evaluate the capability of the Context and Cognitive State triggered Feed-Forward (C2F2) intelligent tutoring system, built upon *AttentiveLearner*, on improving student engagement and learning efficacy in Massive Open Online Course (MOOCs) using real-time bio-signals. When the system detects lack of engagement on the student, it intervenes to remind the student of important content. The authors highlight the importance of keeping the learner engaged to avoid low course completion rates in MOOCs.

The study collected data from n = 48 (20 females) undergraduate and graduate students with an average age of 23.4 years. The researchers collected camera-based photoplethysmography (PPG) using the built-in camera of a Nexus 5 smartphone and a NeuroSky Mindwave EEG headset. The intervention consisted of four phases. First, the introduction phase consisted of a demographic's questionnaire and a 40-second video to introduce the interface. Second, an 18-question content (i.e., computer and network security) pre-test was applied. Third, for the evaluation phase, the students were randomly assigned into four conditions: (1) no feed-forward; (2) context only feed-forward; (3) cognitive only feed-forward; and (4) C2F2. Fourth, in their assigned condition, the learners watched short videos on the learning topics, and, after each video, the participant evaluated it using the Subjective Impression Questionnaire and a content test to measure their understanding on the topic. Finally, the participants completed the Subjective Impression Questionnaire to evaluate the entire lesson and a survey to evaluate the MOOC.

Results from ANOVA analyses to determine the effect of different levels of feed-forward interventions on learning performance showed no significant differences between conditions. Main effect results suggest that the C2F2 model was useful for learners who became disengaged from learning and struggled to refocus. Additionally, a Ranking SVM model was trained to predict participants levels of engagement using EEG-based data from a video feed that achieved a 55.56% accuracy at identifying the low engagement. A second model using PPG-based features was trained to compare the previous model which achieved a 69.44% accuracy, thus outperforming the model using EEG data.

Yang et al. (2019) evaluated an affect detector using logistic regression on a large-scale affect dataset (n = 3,109 observations) from real-life students interacting with the ASSISTments ITS. Affective states are highly predictive of students' academic outcome, thus creating an environment where the student stays in a positive affective state, such as engaged, can be greatly beneficial to obtaining higher academic outcomes. The authors emphasize that the purpose of the study was to determine if a simple model, like the one proposed, could collect better data than more refined and computer expensive models.

The dataset has been widely used for training other affect sensors. Each instance consists of a label of the affective state detected in a window of time of 20 seconds and features to identify the task the student interacted with from the ASSISTments (i.e., time the student spent in the task, number of hints the student asked for, etc.). However, the dataset was reduced to the instances where the students were labeled into any of the four BROMP affective states: boredom, confusion, engaged concentration, and frustration. The Linear Minimum Mean-Square Error (L-MMSE)-based model was compared to three more active learning methods: uncertainty sampling (US), expected variance reduction (EVR), and model change (MC). Two settings of cross validation were applied to the model: (1) treating each of the instances as independent students and (2) merging every instance of a student and treating it as one instance.

Two models were created: the first one to detect engaged concentration, due to being the most recurrent affective state of the cleaned dataset (82%). Results from the AUC, the capability of a model to distinguish between classes, indicate that the proposed model outperformed the other models at classifying between both cross-validation settings. The second model was created to the other three affective states from the dataset (bored, confused, and frustrated). Due to the instances of these states being rare, resampling was used to balance the dataset. Results from the AUC indicate that the proposed model outperformed the other models at classifying between both cross-validation settings. Although the performance metric is significantly smaller compared to the model to detect engaged concentration.

In sum, the proposed active learning methods are effective at making informative observations, thus reducing the number of instances needed for a model to accurately detect affective states. One of the limitations of the study is that the proposed model is only able to perform binary classifications between affective states (i.e., engaged or not engaged). However, the authors point out that creating multi-class classification models are more useful for classroom settings.

Yue et al. (2019) conducted a quasi-experimental study to propose a framework to detect students' learning engagement when interacting with e-learning platforms using multimodal data, such as video feed from a webcam, eye tracking information, and clickstream data from a MOOC to learn Python programming. Data for this study was acquired from two databases used to train facial expressions: ImageNet, used to pre-train the models, and USTC-NVIE, used for the model to learn Asian face features. The researchers developed an eye learning behavioral database using a Tobii Eye Tracker 4C in n = 22 subjects.

The study collected data from n = 46 (14 females) undergraduate (n = 33), graduate (n = 11), and doctoral (n = 2) students. Majority of the subjects (66%) were enrolled in a Computer Science and Technology program. Before the intervention, the students answered a demographics survey that asked for their prior Python language knowledge level. After watching a short video on the MOOC, the learners self-assessed their learning performance from 10 to 100. Finally, they answered a knowledge post-test. The intervention lasted on average 50 minutes per participant.

To evaluate the proposed facial expression model, four models were compared: (1) VGG16 (Convolutional Neural Network with 16 layers) without Long Short-Term Memory (LSTM), (2) Inception-ResNetV2 without LSTM, (3) VGG16 with LSTM, and (4) Inception-ResNetV2 with LSTM. The LSTM models were the most accurate, meaning that a combination of spatial and temporal features can improve models for facial expression classification. The best performing model was the VGG16 with LSTM (76.08% accuracy). To evaluate the proposed eye behavior model, three models were compared: (1) Classification and Regression Trees (CART), Random Forest, and Gradient Boosted Decision Tree (GBDT). The GBDT was the best performing algorithm with an 81% accuracy.

In sum, the proposed framework was proven to be effective. However, one limitation of the study was that it only focused on distance-learning environments from PC settings and not mobile devices.

Tables

Table A1

Table of Reviewed Publications

Publication ID	Citation in APA Format
1	Barron-Estrada, M. L., Zatarain-Cabada, R., & Aispuro-Gallegos, C. G. (2018). Multimodal recognition of emotions with application to mobile learning. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 416–418. https://doi.org/10.1109/ICALT.2018.00104
2	Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 379–388. https://doi.org/10.1145/2678025.2701397
3	Burleson, W., & Picard, R. W. (2007). Gender-Specific Approaches to Developing Emotionally Intelligent Learning Companions. *IEEE Intelligent Systems*, *22*(4), 62–69. https://doi.org/10.1109/MIS.2007.69
4	DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193. https://doi.org/10.1007/s40593-017-0152-1
5	Ghaleb, E., Popa, M., Hortal, E., Asteriadis, S., & Weiss, G. (2018). Towards affect recognition through interactions with learning materials. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 372–379. https://doi.org/10.1109/ICMLA.2018.00062
6	Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. *Proceedings of the 16th International Conference on Multimodal Interaction*, 42–49. https://doi.org/10.1145/2663204.2663264
7	Joshi, A., Allessio, D., Magee, J., Whitehill, J., Arroyo, I., Woolf, B., Sclaroff, S., & Betke, M. (2019). Affect-driven Learning Outcomes Prediction in Intelligent Tutoring Systems. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–5. https://doi.org/10.1109/FG.2019.8756624
8	Jraidi, I., & Frasson, C. (2013). Student's uncertainty modeling through a multimodal sensor-based approach. *Journal of Educational Technology & Society*, *16*(1), 219–230.
9	Jraidi, I., Chaouachi, M., & Frasson, C. (2013). A dynamic multimodal approach for assessing learners' interaction experience. *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, 271–278. https://doi.org/10.1145/2522848.2522896
10	Kai, S., Paquette, L., Baker, R. S., Bosch, N., D'Mello, S., Ocumpaugh, J., Shute, V., & Ventura, M. (2015). A comparison of video-based and interaction-based affect detectors in Physics Playground. In *International Educational Data Mining Society*. http://eric.ed.gov/?id=ED560544
11	Korn, O., & Rees, A. (2019). Affective effects of gamification: Using biosignals to measure the effects on working and learning users. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 1–10. https://doi.org/10.1145/3316782.3316783

12	Lee-Cultura, S., Sharma, K., Papavlasopoulou, S., Retalis, S., & Giannakos, M. (2020). Using sensing technologies to explain children's self-representation in motion-based educational games. *Proceedings of the Interaction Design and Children Conference*, 541–555. https://doi.org/10.1145/3392063.3394419
13	Leong, F. H. (2015). Automatic detection of frustration of novice programmers from contextual and keystroke logs. *2015 10th International Conference on Computer Science & Education (ICCSE)*, 373–377. https://doi.org/10.1109/ICCSE.2015.7250273
14	Mangaroska, K., Sharma, K., Gaševic, D., & Giannakos, M. (2020). Multimodal learning analytics to inform learning design: Lessons learned from computing education. *Journal of Learning Analytics*, *7*(3), 79–97.
15	Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Put your thinking cap on: Detecting cognitive load using EEG during learning. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 80–89. https://doi.org/10.1145/3027385.3027431
16	Muñoz, K., Noguez, J., Neri, L., Kevitt, P. M., & Lunney, T. (2016). A computational model of learners achievement emotions using control-value theory. *Journal of Educational Technology & Society*, *19*(2), 42–56.
17	Park, J., Park, Y. H., Kim, J., Cha, J., Kim, S., & Oh, A. (2018). Elicast: Embedding interactive exercises in instructional programming screencasts. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10. https://doi.org/10.1145/3231644.3231657
18	Pham, P., & Wang, J. (2016). Adaptive review for mobile MOOC learning via implicit physiological signal sensing. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 37–44. https://doi.org/10.1145/2993148.2993197
19	Psaltis, A., Apostolakis, K., Dimitropoulos, K., & Daras, P. (2017). Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Computational Intelligence and AI in Games*, *10*, 1–1. https://doi.org/10.1109/TCIAIG.2017.2743341
20	Rajendran, R., Iyer, S., & Murthy, S. (2019). Personalized affective feedback to address students' frustration in ITS. *IEEE Transactions on Learning Technologies*, *12*(1), 87–97. https://doi.org/10.1109/TLT.2018.2807447
21	Sharma, K., Papavlasopoulou, S., & Giannakos, M. (2018). Coding games and robots to enhance computational thinking: How collaboration and engagement moderate children's attitudes? *International Journal of Child-Computer Interaction*, *21*, 65–76. https://doi.org/10.1016/j.ijcci.2019.04.004
22	Sharma, K., Papavlasopoulou, S., Lee-Cultura, S., & Giannakos, M. (2021). Information flow and children's emotions during collaborative coding: A causal analysis. *Interaction Design and Children*, 350–362. https://doi.org/10.1145/3459990.3460731
23	Sinha, A., Gavas, R., Chatterjee, D., Das, R., & Sinharay, A. (2015). Dynamic assessment of learners' mental state for an improved learning experience. *2015 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/FIE.2015.7344121
24	Sottilare, R. A., & Proctor, M. (2012). Passively classifying student mood and performance within intelligent tutors. *Journal of Educational Technology & Society*, *15*(2), 101–114.
25	Srivastava, N., Newn, J., & Velloso, E. (2018). Combining low and mid-level gaze features for desktop activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(4), 189:1-189:27. https://doi.org/10.1145/3287067

26	Standen, P. J., Brown, D. J., Taheri, M., Galvez Trigo, M. J., Boulton, H., Burton, A., Hallewell, M. J., Lathe, J. G., Shopland, N., Blanco Gonzalez, M. A., Kwiatkowska, G. M., Milli, E., Cobello, S., Mazzucato, A., Traversi, M., & Hortal, E. (2020). An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *British Journal of Educational Technology*, *51*(5), 1748–1765. https://doi.org/10.1111/bjet.13010
27	Subburaj, S. K., Stewart, A. E. B., Ramesh Rao, A., & D'Mello, S. K. (2020). Multimodal, multiparty modeling of collaborative problem solving performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 423–432). Association for Computing Machinery. http://doi.org/10.1145/3382507.3418877
28	The, B., & Mavrikis, M. (2016). A study on eye fixation patterns of students in higher education using an online learning system. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 408–416. https://doi.org/10.1145/2883851.2883871
29	Vail, A. K., Grafsgaard, J. F., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2016). Gender differences in facial expressions of affect during learning. *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 65–73. https://doi.org/10.1145/2930238.2930257
30	VanLehn, K., Zhang, L., Burleson, W., Girard, S., & Hidago-Pontet, Y. (2017). Can a non-cognitive learning companion increase the effectiveness of a meta-cognitive learning strategy? *IEEE Transactions on Learning Technologies*, *10*(3), 277–289. https://doi.org/10.1109/TLT.2016.2594775
31	Veliyath, N., De, P., Allen, A. A., Hodges, C. B., & Mitra, A. (2019). Modeling students' attention in the classroom using eyetrackers. *Proceedings of the 2019 ACM Southeast Conference*, 2–9. https://doi.org/10.1145/3299815.3314424
32	Wu, C.-H., Tzeng, Y.-L., & Huang, Y.-M. (2020). Measuring performance in leaning process of digital game-based learning and static E-learning. *Educational Technology Research and Development*, *68*(5), 2215–2237. https://doi.org/10.1007/s11423-020-09765-6
33	Xiao, X., & Wang, J. (2015). Towards attentive, bi-directional MOOC learning on mobile devices. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 163–170. https://doi.org/10.1145/2818346.2820754
34	Xiao, X., & Wang, J. (2016). Context and cognitive state triggered interventions for mobile MOOC learning. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 378–385. https://doi.org/10.1145/2993148.2993177
35	Yang, T.-Y., Baker, R. S., Studer, C., Heffernan, N., & Lan, A. S. (2019). Active learning for student affect detection. *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019. International Educational Data Mining Society (IEDMS) 2019*, 208–217.
36	Yue, J., Tian, F., Chao, K.-M., Shah, N., Li, L., Chen, Y., & Zheng, Q. (2019). Recognizing multidimensional engagement of e-learners based on multi-channel data in e-learning environment. *IEEE Access*, *7*, 149554–149567. https://doi.org/10.1109/ACCESS.2019.2947091