Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 09/Apr/2024; Camera ready: 07/Oct/2024; 1st round notif.: 21/Sep/2024; Edition review: 21/Nov/2024; Available online: 21/Nov/2024; Published: 21/Nov/2024;

Sensor-free Affect Detection in Learning Environments: A Systematic Literature Review

Felipe de Morais
UNISINOS
ORCID: 0000-0002-8510-4516
felipmorais@unisinos.br

Diógines Goldoni
UNISINOS
ORCID: 0000-0001-8675-6287
davilag@unisinos.br

Tiago R. Kautzmann
Instituto Federal de Educação do Paraná (IFPR)
ORCID: 0000-0002-6017-8340
tiago.kautzmann@ifpr.edu.br

Patricia A. Jaques
PPGC/UFPEL; PPGInf/UFPR
ORCID: 0000-0002-2933-1052
patricia.jaques@inf.ufpel.edu.br; patricia@inf.ufpr.br

Abstract

Emotions and affective states influence cognition and learning processes. Computer-based learning environments (CBLEs) capable of detecting and adapting to these states significantly enhance learning outcomes. However, practical constraints often hinder the deployment of sensor-based affect detection in CBLEs, especially for large-scale or long-term use. Consequently, sensor-free affect detection, reliant solely on interaction logs, emerges as a promising alternative. This paper offers a comprehensive literature review on sensor-free affect detection, covering frequently identified affective states, methodologies for sensor development, CBLE attributes, and research trends. Despite the field's maturity, there's ample room for further exploration. Future research should focus on improving sensor-free detection models, collecting more samples of underrepresented emotions, and refining model development practices. Additionally, efforts should be made to integrate models into CBLEs for real-time detection, provide meaningful interventions based on detected emotions, and deepen understanding of emotions' impact on learning. Key suggestions include comparing data collection techniques, optimizing duration granularity, establishing shared databases, and ensuring model source code accessibility.

Keywords: *Sensor-Free Affect Detection; Systematic Literature Review; Emotional Learning Environments; Emotion Detection; Computer-Based Learning Environments.*

Cite as: Morais, F., Goldoni, D., Kautzmann, T. R. Jaques, P. A. (2024). Sensor-free Affect Detection in Learning Environments: A Systematic Literature Review. Revista Brasileira de Informática na Educação, 32, 679-717. https://doi.org/10.5753/rbie.2024.4362.

1 Introduction

Various cognitive and affective factors play pivotal roles in shaping student learning processes. A blend of elements such as affect, motivation, and metacognition interacts with cognition to influence learning outcomes (Azevedo & Aleven, 2013). Among these factors, emotions and affective states stand out as particularly important in determining the success or difficulty of learning. Emotions influence a student's attention, motivation, cognition, and self-regulation, affecting how learning progresses (Pekrun et al., 2014). Positive emotions, such as joy and engagement, can foster intellectual and creative development (Fredrickson, 1998), while negative emotions, like frustration and boredom, may disrupt the learning process (Graesser & D'Mello, 2011). Confusion, depending on how it is handled, can either hinder or improve learning (D'Mello et al., 2014).

As research advances our understanding of the role of emotions in learning, its application in Computer-Based Learning Environments (CBLEs) becomes essential. Studies show that CBLEs capable of detecting and adapting to students' affective states can significantly improve learning outcomes (Arroyo et al., 2016; S. D'Mello et al., 2010; Litman & Forbes-Riley, 2014). However, achieving this "affect-aware" adaptation requires the accurate detection and interpretation of emotions, which remains a significant challenge in the field.

Traditional affect detection methods, such as analyzing facial expressions, vocal tones, or body posture, often rely on costly and intrusive technologies like cameras or specialized sensors (Calvo & D'Mello, 2010). While effective, these methods are not practical for large-scale or long-term use in learning environments (Baker et al., 2012). In contrast, sensor-free affect detection refers to the detection of emotions without the use of external hardware, such as cameras or wearable sensors. Instead, this approach relies on input devices like keyboards and mice—commonly used as part of the learning environment itself—to track students' interactions and infer emotional states (Paquette et al., 2016; Salmeron-Majadas et al., 2014). This method avoids the need for additional equipment by mining interaction logs from the graphical user interface (GUI) (Baker et al., 2012), thus making it more scalable and non-intrusive for students.

Sensor-free affect detection leverages machine learning techniques to analyze the data produced by students' interactions. The process begins with the collection of interaction logs, paired with concurrent emotional annotations. After cleaning and preprocessing the data, it is used to train machine learning models to recognize patterns associated with specific emotions. These models can then detect emotions passively, offering a non-intrusive means of tracking affect in real time without the need for additional sensor devices.

This method presents multiple advantages. First, it allows for the large-scale monitoring of students' emotions in a non-intrusive manner, which is critical for personalized learning. Second, sensor-free systems can be implemented at scale, making them ideal for online learning platforms. Finally, the passive nature of this data collection reduces the burden on both students and educators, allowing for real-time feedback and timely interventions when students struggle.

While significant progress has been made in this area, recent advancements have yet to be comprehensively reviewed. The most notable overview of sensor-free affect detection is provided by Baker and Ocumpaugh (2014), who offer a thorough exploration of early works in the field. However, their review lacks a systematic approach and does not include the latest developments from the past decade.

Thus, this paper aims to fill this gap by presenting a systematic review of the recent literature on sensor-free affect detection in learning environments. Our review addresses key research questions related to current trends, methodologies, and outcomes in this area, providing a more up-to-date and structured understanding of the field.

2 Affect in learning

This article investigates affect detection within CBLEs. To provide a foundation, it is important to first define the constructs of "affect", "affective states", and "emotions" that form the basis of our exploration. We adopt Scherer's classification, as outlined in (Scherer, 2000), which considers emotions as one aspect of affective states, accompanied by moods and dispositions such as personality traits. Emotions, characterized by their relatively brief duration and high intensity, are typically elicited by specific incidents. According to Scherer, emotions are merely one type of affective state, with affective states encompassing a broader spectrum of emotional manifestations.

Scherer later introduced the term "affective phenomena" instead of "affective state", emphasizing the dynamic and process-oriented nature of these phenomena across multiple components (Scherer, 2005). Consequently, the term "affect" is used interchangeably with "affective phenomena". This concept aligns with (Russell, 2003)'s description of "core affect" as a "neurophysiological state consciously accessible as a straightforward, non-reflective feeling comprising an integral mix of hedonic (pleasure–displeasure) and arousal (sleepy–activated) values".

The duration of emotions varies widely depending on the nature of the emotion, the context in which it occurs, and individual characteristics (Verduyn et al., 2011). Basic emotions such as joy, sadness, anger, and fear can last from seconds to minutes, but when rumination occurs—repeatedly thinking about the event that triggered the emotion—these emotions can extend for hours or even days (Verduyn & Lavrijsen, 2015). Factors such as chronic stress or psychological conditions like depression can further prolong emotions like sadness and anxiety. Additionally, the intensity and meaning attributed to the emotion influence its duration, and how people regulate their emotions is crucial in determining how long they persist (Verduyn et al., 2015).

Importantly, not all affective phenomena associated with learning are classified as emotions. For instance, interest, while crucial to the learning experience, is often categorized as a motivational state rather than a conventional emotion (Hidi, 2006). Given these nuances and following the approach of scholars in the field of sensor-free affect detection (Andres et al., 2019; Baker et al., 2014; Botelho et al., 2017; Pardos et al., 2014), we adopt the term "affect" as synonymous with Scherer's "affective phenomena," which includes emotions and other affective phenomena.

Although researchers in sensor-free emotion detection typically aim to discern affect, emotions have received more extensive scrutiny in learning settings. Emotions commonly observed in learning environments are referred to as learning-centered emotions (Graesser & D'Mello, 2012; Graesser et al., 2014) or academic emotions, as per the Control-Value Theory (CVT) (Pekrun et al., 2002b, 2006). Academic emotions can be categorized into groups such as achievement emotions, epistemic emotions, and others (e.g., social emotions like anger) (Pekrun & Linnenbrink-Garcia, 2012). Achievement emotions are related to learning activities (e.g., frustration, boredom) or their outcomes (e.g., pride, anxiety).

In conclusion, the Control-Value Theory (CVT) suggests that emotions are shaped by two key types of appraisals. First, emotions are influenced by the degree of control students perceive they have over their learning activities and the outcomes of those activities. Second, emotions are shaped by the subjective value students assign to these activities and outcomes, that is, how important or relevant they consider them. Additionally, epistemic emotions—such as confusion, surprise, and interest—are often triggered by cognitive factors, including the way students process task-related information. Emotions can be further categorized along two dimensions: valence (whether the emotion is positive or negative) and arousal (whether the emotion is activating or deactivating). These dimensions combine to form four emotional categories: positive activating (e.g., enjoyment and curiosity/interest), positive deactivating (e.g., relaxation), negative activating (e.g., anger, frustration, confusion, and anxiety), and negative deactivating (e.g., boredom).

3 Sensor-free Affect Detection

This study aims to understand sensor-free detectors, which employ data derived solely from student interactions with the CBLE through input devices such as keyboards and mouse. These interactions may include a variety of actions such as keystroke count, typing speed, and mouse movements (Vea, Rodrigo, et al., 2016), and can extend to information extracted from the CBLE, like the number of errors a student commits on a task (Wang et al., 2015), the frequency of help requests (Baker et al., 2012), and achieved goals (Sabourin et al., 2011). These models, termed as “sensor-free detectors,” have exhibited promising results in the literature (Arroyo & et al., 2009; Baker et al., 2012; Botelho et al., 2017; Paquette et al., 2016).

Primarily, sensor-free detectors are developed as classifier or regression models, utilizing a supervised learning approach. Their development process typically encompasses five main phases: data collection, feature engineering, model development, performance analysis, and application, as depicted in Figure 1. Although this method is broadly applied across sensor-free affect detector development, the execution of each phase is contingent on the specific research objectives.

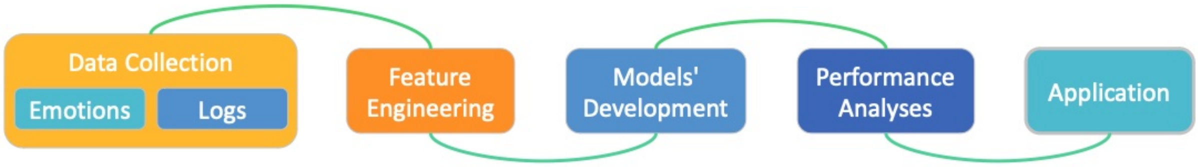

Figure 1: Phases of development of sensor-free affect detectors.

- **Data Collection:** This phase focuses on gathering the requisite information for the development of sensor-free detection models. It necessitates two types of data: action logs and emotions, which we categorize as sub-phases.
  - **Logs:** Logs encapsulate student actions during CBLE interactions and are the models input. These actions can be varied, from clicks on the Graphical User Interface (GUI)

to task completion, mistakes made, feedback received, and duration of each action, among others. The nature of the collected logs is determined by the specific CBLE in use.

- **Emotions:** Emotions serve as ground truth labels (output) for supervised machine learning training, crucial for the development of sensor-free affect detectors. These labels, synchronized with log data and calibrated to the observation window duration, can be collected using either a first-person or third-person approach or a mixed one. The first-person approach utilizes self-reporting, where students express their own emotions. This is typically achieved through periodic questionnaires, self-report scales (Wixon & Arroyo, 2014), or 'emote-aloud' protocols (Porayska-Pomsta et al., 2013), either administered at regular intervals or at the conclusion of a learning session. While providing direct access to a student's perceived emotions, periodic questionnaires can be interruptive and disruptive to the learning process. Additionally, due to social desirability bias or limited emotional self-awareness, students may not always accurately self-report their emotions. Contrarily, the third-person approach incorporates observer ratings, wherein trained observers, either co-located with students (Pedro et al., 2013) or viewing recorded student videos (de Morais & Jaques, 2023), label emotional states based on exhibited behavior. This method, potentially less intrusive, can offer a more objective perspective of a student's emotional state. Nevertheless, it may not always accurately capture the student's internal emotional state, particularly when certain emotions are not clearly manifested in observable behavior. Moreover, this method necessitates observer training.

These sub-phases may occur concurrently or sequentially. For methods that rely on crowdsourcing and log file annotation, log data collection is a prerequisite since it informs the analysis employed in emotion detection. Conversely, methods based on human observations or student self-reports can transpire alongside log collection. In essence, the synchronization between logs and emotions is a function of the approach adopted and refers to the accurate temporal and sequential identification of the emotion a student experiences when generating system interaction logs.

- **Feature Engineering:** After the data collection phase, an essential step in developing sensor-free affect detectors is feature engineering or data pre-processing. Feature engineering involves the transformation of raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy. In the context of sensor-free affect detectors, feature engineering might involve creating new variables such as the average time between actions, the number of errors per minute, or the frequency of specific action sequences. For example, Botelho et al. (2019) generated 92 features to detect affect and off-task behavior. These features describe actions over a 20-second window, capturing both immediate behavior and the broader context of the interactions. These variables were used to train models such as Naive Bayes and LSTM, improving the ability to detect student behaviors and affective states.

- **Model Development:** This phase comprises the selection of appropriate machine learning algorithms and the creation of emotion detection models through training. These models are trained with labeled data, i.e., data generated from logs and synchronized with emotion labels. Generally, in the realm of sensor-free affect detection, separate models are developed for each emotion to identify the best model for each emotion independently. The phase also includes data normalization, handling of missing data, class balancing, and feature selection for log representation. The model development process typically begins with the selection of machine learning algorithms, such as decision trees, neural networks, support vector machines, or ensemble-based algorithms like random forests. These algorithms are evaluated based on their suitability for detecting specific emotions, and multiple models are often tested to determine which provides the best results. Once an algorithm is selected, the training of models is conducted using interaction data from students, synchronized with emotion labels. Cross-validation techniques are applied to ensure that the model generalizes well to new data, preventing overfitting. Additionally, hyperparameter tuning is performed, often through grid search or Bayesian optimization, to refine the model's parameters (e.g., learning rate, tree depth) and improve performance. During this phase, it is also essential to address issues such as class balancing because certain emotions, like boredom or confusion, may be underrepresented in the data. Techniques like oversampling, undersampling, or SMOTE (Synthetic Minority Over-sampling Technique) can be employed to manage class imbalance. Furthermore, feature selection plays a critical role in improving the accuracy of the model by selecting the most relevant features from the logs, such as the number of actions, task completion rates, or frequency of help requests. This step is crucial for reducing noise and ensuring that the model focuses on the most informative features.

- **Performance Analysis:** This phase aims to evaluate the efficacy of the developed models. Typically, model performance is gauged using various metrics that, in essence, indicate the likelihood of accurate student emotion classification based solely on log data. To assess model performance, unseen data is utilized for testing. Various validation methods are also applied to ensure the generalizability of model performance when deployed on different samples. Models are evaluated based on performance metrics such as AUC, F1-Score, precision, and Cohen's Kappa, to assess their ability to correctly predict emotions. These metrics help not only in identifying accuracy but also in understanding the quality of classification, particularly in cases of class imbalance, such as with less frequent emotions like frustration or confusion. For instance, in the study by (Paquette et al., 2014), the performance of the emotion detection model was assessed using the Cohen's Kappa metric, achieving a average value of 0.354 for detecting frustration, boredom, confusion and concentration. This indicates moderate agreement between the emotions detected by the model and the emotions labeled by human observers, demonstrating that the model could correctly identify emotions in learning activities, although there was still room for improvement in accuracy.

- **Application:** This phase entails the utilization of developed models in experiments, analyses, or comparisons. For example, (Paquette et al., 2016) successfully integrated a sensor-free affect detector into the vMedic - TC3Sim, a medical training simulation. The model, developed to detect emotions like confusion and frustration, was applied in real-time during the training simulation to provide feedback to medical trainees. This implementation demonstrated the potential for affect-aware learning environments to adaptively respond to learners' emotional states, improving the overall learning experience.

As demonstrated, the development of sensor-free affect detection models demands a high level of methodological rigor at every stage, from data collection to model deployment. Accurate and synchronized data collection, whether through student interaction logs or emotion labeling, is crucial for producing reliable models. Equally important is the thoughtful design of feature engineering and model development processes, ensuring that features are well-aligned with the emotional phenomena being studied. Given the inherent variability in student behavior and emotions, ensuring the integrity of the data and properly handling issues such as class imbalance and missing data are essential for building robust and generalizable models. By maintaining a focus on the quality of data collection and processing, sensor-free affect detection systems can provide valuable insights into student emotions, ultimately improving personalized learning experiences across diverse educational contexts.

4 Method

This paper provides an overview of current knowledge on detecting emotions based solely on students' actions within Computer-Based Learning Environments (CBLEs). To achieve this, we employed a systematic review method following the five-step process proposed by Petersen et al. (2008): *i*) define the research questions, *ii*) perform the search for primary studies, *iii*) screen the papers, *iv*) apply keywords to abstracts, and *v*) extract and map the data.

To guide the systematic review, we formulated four Research Questions (RQs):

- **RQ1:** Which emotions are being detected by sensor-free detectors in CBLEs, and upon what theoretical foundations are these detections based?
- **RQ2:** What are the methodologies and technologies behind the development of sensor-free affect detectors specifically for CBLEs?
- **RQ3:** What are the primary contexts in which sensor-free emotion detectors are employed within CBLEs?
- **RQ4:** How do sensor-free emotion detectors perform in terms of generalization and practical application within CBLEs, and what does the current research landscape reveal about trends, key findings, and future directions?

4.1 Search procedures

We started by examining a number of initial papers provided by experts in the field. Then, we defined our research questions, and based on the keywords from the RQs, we created the search string. The search was conducted in the most relevant digital libraries that index papers on Computer Science, Affective Computing, Educational Data Mining, and Learning Analytics: Web of Science, IEEE Xplore, Science Direct, ACM Digital Library, Engineering Village, ERIC, JEDM, Scopus, and Springer. Different search strings were used according to the rules and syntax of each digital library. However, the core was based on *"(sensor-free OR interaction-based OR log-based) AND (emotion OR affect) AND (detect OR predict OR infer)"*. Although the search string does not explicitly include terms such as 'learning' or 'learning environments,' it was designed to focus on emotion detection approaches (sensor-free or log-based) that are frequently applied in Computer-Based Learning Environments (CBLEs). We assumed that works in this domain typically mention or imply a learning context, even without directly using these terms, due to the prominence of educational settings in research on affect detection. Furthermore, the learning context was implicitly addressed during the screening and filtering processes, ensuring that only studies relevant to CBLEs were included. The primary works were searched between November and December of 2021[1].

We applied inclusion and exclusion criteria to filter the primary works. The Exclusion Criteria (EC) were: (EC1) Duplicated papers; (EC2) Papers not written in English; (EC3) Papers published as secondary or tertiary studies; (EC4) Non-peer-reviewed papers, books, proceedings, dissertations, thesis, summary, or short papers; (EC5) Papers focusing solely on affect detection based on text (i. e., sentiment analysis); (EC6) Papers focusing on sensors for detecting affect [2]; (EC7) Papers focusing on affect detection in areas other than education; (EC8) Papers focusing on different areas than automatic affect detection. The Inclusion Criteria (IC) were: (IC1) Studies that detect emotions through logs generated by the student's interaction with the learning environment; (IC2) Studies comparing data-driven and sensor-driven approaches, that also provide details about the development of the models; (IC3) If several papers reported the same study, only the most recent was included.

To ensure the relevance and accuracy of the findings, future studies could consider the inclusion of a temporal cutoff. For instance, restricting the analysis to studies published within the last ten years could mitigate the inclusion of potentially outdated research while maintaining a focus on current methodologies and technologies. However, in this review, we chose not to impose such a temporal restriction to capture the full development of sensor-free affect detection models, including foundational studies that continue to inform current research in the field.

The filtering process, guided by the application of the exclusion and inclusion criteria, is depicted in Figure 2. Our initial search yielded 1,039 papers. Of these, 225 were removed due to duplication. The remaining 814 papers were randomly assigned to three reviewers (the authors) for initial screening based on the title and keywords, resulting in the rejection of 658 papers. The reviewers then assessed the abstracts of the remaining 156 papers, eliminating another 77. The full texts of the remaining 79 articles were screened, with 46 more papers being excluded due to insufficient information about the development of the emotion detection models. The reviewers conducted a thorough review of the final 33 articles, resulting in the elimination of 14 more. Specifically, 3 papers were excluded as they were non-peer-reviewed papers, books, proceedings, dissertations, thesis, summaries, or short papers (EC4), 3 papers were excluded for focusing on sensor-based affect detection (EC6), 1 paper was excluded for focusing on affect detection outside the realm of education (EC7), and 7 papers were excluded for focusing on areas other than automatic affect detection (EC8). Ultimately, 19 papers were included in the final selection. The complete list of papers at each phase, along with the corresponding criteria, can be found in the external worksheet at https://bit.ly/44nJGhC.

[1] The time elapsed between the search and the publication of this article is due to the extended duration of the systematic review process, which took over a year to complete, as well as the additional time spent in peer review and publication. We believe the findings remain representative of the field's state during the review period.

[2] This criterion includes sensors in mobile phones, for example, touchscreen pressure data, camera, etc.

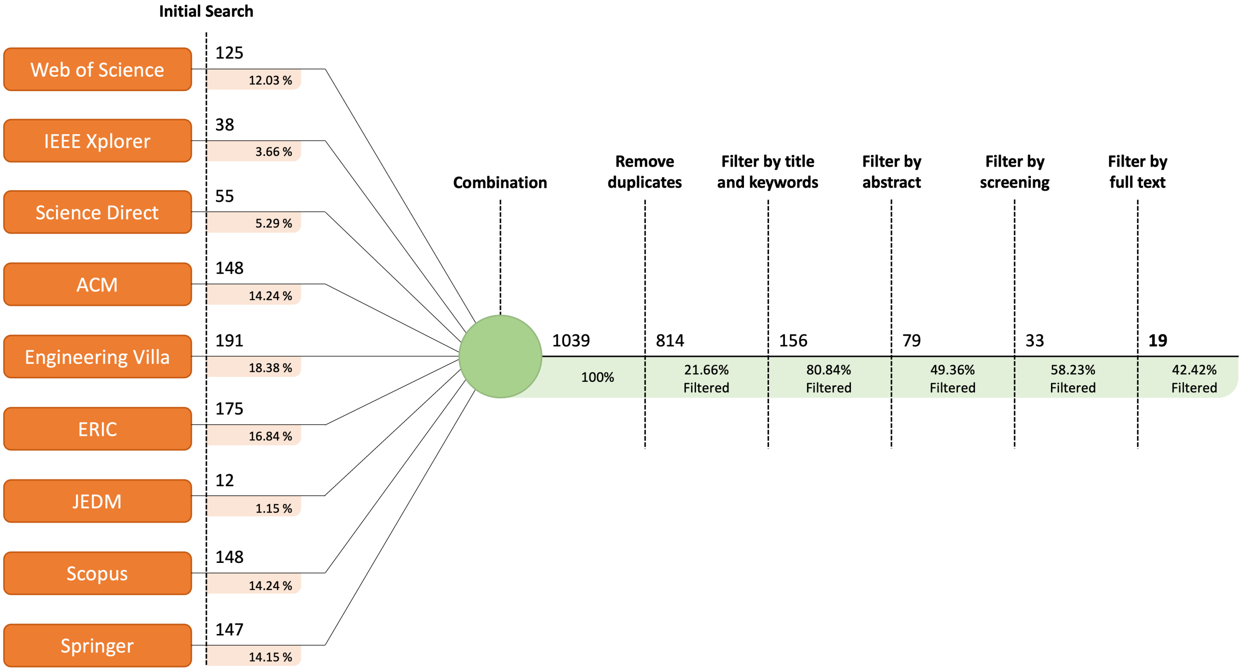

Figure 2: Papers selection process according to each digital library and the phases.

5 Results and Discussions

In this section, we delve into the findings and engage in discussions centered around the previously outlined research questions (RQs). Our exploration is systematically structured to dissect each RQ through its sub-questions, aiming to provide a comprehensive understanding of the distinct facets within sensor-free detection of emotions in CBLEs.

5.1 RQ1 - Which emotions are being detected by sensor-free detectors in CBLEs, and upon what theoretical foundations are these detections based?

To thoroughly investigate the spectrum of emotions identified by sensor-free detectors in CBLEs and the theoretical underpinnings guiding these detections, RQ1 has been dissected into five sub-questions. Each sub-question is designed to explore a distinct dimension of emotion detection within CBLEs: from the specific emotions considered and their methods of collection, to the granularity of detection, the theoretical frameworks employed, and the inclusion of supplementary information. This structured approach allows for a comprehensive analysis of how emotions are detected, understood, and interpreted in CBLEs.

RQ1.1 - What specific emotions are sensor-free detectors in CBLEs designed to recognize?

To address this question, we compiled a list of all the emotions that emotion detectors in the selected works consider and recorded the frequency of each detected emotion. Figure 3 shows that boredom, confusion, frustration, and engagement are the most detected emotions in more than 84% of the works. A possible justification for almost all studies considering these emotions in detection models is that these emotions are found to be more frequent in complex learning activities (S. D'Mello & Graesser, 2012) and during learning with technologies (Bosch & D'Mello, 2017; Bosch et al., 2013; S. D'Mello, 2013).

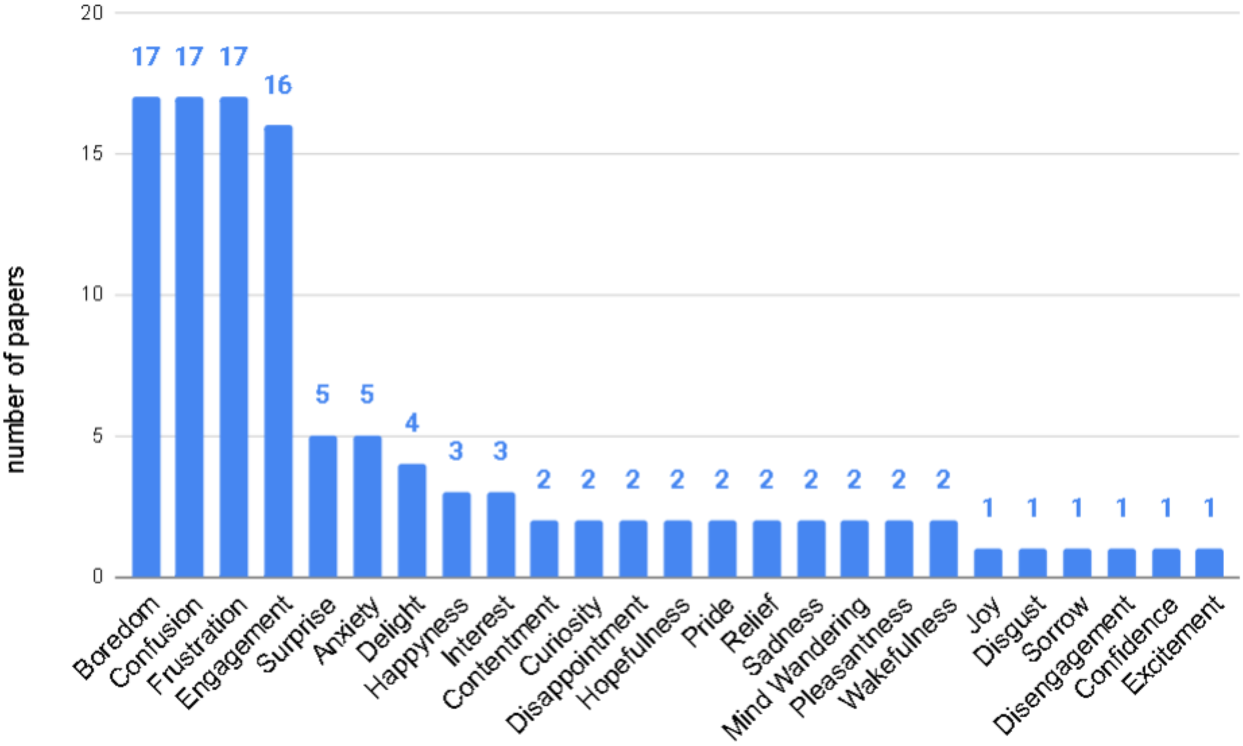

Figure 3: Emotions considered by the emotion detectors.

RQ1.2 - What methods are employed to capture these emotions in CBLEs?

The selected works employed various methods to capture students' emotions. These methods can be categorized into third-person and first-person approaches. The third-person approach includes annotations from human observers, crowdsourcing, and log-file annotation, while the first-person approach involves student self-reports.

Annotations from human observers, which is the most commonly used method, were utilized by several works (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2014, 2016; Pardos et al., 2014). In this type of data collection, experts known as "coders" (typically trained professionals or researchers with experience in behavioral analysis and affect detection) observe students and annotate their emotions by analyzing their facial and physical expressions during the learning activity. This observation can be conducted through videos (offline) or directly in the classroom (online). Although other human-observer-based methods exist, such as the EmAP-ML protocol (de Morais et al., 2019), the most widely referenced method in this category is the Baker-Rodrigo-Ocumpaugh Method and Protocol (BROMP), a protocol used for annotating students' emotions and behaviors in real-time during their interactions with a CBLE (Ocumpaugh et al., 2015)

The second most commonly used method, student self-reports, was employed by (Jensen et al., 2019), (Wixon & Arroyo, 2014), (Hutt et al., 2019), (Smeets et al., 2019), and (Tiam-Lee & Sumi, 2019). This method requires students to report their emotions during learning activities. For instance, the emote-aloud method (Porayska-Pomsta et al., 2013) involves students verbally expressing their emotions while interacting with the CBLE. Another approach is the free-response method (S. K. D'Mello et al., 2006), in which students use specific affect terms (e.g., happy), valence terms (e.g., slightly confused), or arousal terms (e.g., very active) to express their emotions in their own words.

The other two categories, crowdsourcing and log-file annotation, were each utilized by only one study. (Yang et al., 2016) employed the crowdsourcing method, which involved a large group of paid participants annotating the students' confusion levels while reading their posts from the course forum. In this method, five coders assigned a confusion level to each sample, and the outcomes were determined based on the average or number of votes. On the other hand, log-file annotation, as applied by (Cocea & Weibelzahl, 2010), involved retrospective analysis of student compilation logs in the system. Coders annotated students' emotions based on their perceptions of the data.

RQ1.3 - What is the granularity of emotion detection in these studies?

The aim of this RQ is to determine the duration of the observation window for annotating and capturing students' actions within the system. In the literature, the term "grain level" (or "granularity") has been used to define this time period for obtaining emotions. Based on the examined papers, this time period can vary from seconds to weeks, depending on the approach used to collect emotions.

In the study by (Tiam-Lee & Sumi, 2019), students' self-reports were annotated with variable time intervals, with an average of 17 seconds. Additionally, the collection process allowed the same student to make multiple annotations consecutively. For works that followed the BROMP protocol, each emotion was associated with a 20-second window (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2014, 2016; Pardos et al., 2014). However, BROMP coders analyzed students in a round-robin fashion, meaning that the time between consecutive emotion labels could be minutes apart.

In (Wixon & Arroyo, 2014)'s study, the system prompted students to identify their current emotions, with the pop-up appearing every 5 to 7 minutes. Similarly, (Smeets et al., 2019) requested students to self-report their emotions based on the same strategy, but with a maximum of three pop-ups per week. (Jensen et al., 2019) and (Hutt et al., 2019) also employed a pseudorandomly triggered pop-up strategy based on students' platform activities on their studies which aim was to investigate the generazibility of detection models across different student populations using data from 69,174 students over an entire school year. These studies allowed for one pop-up every two weeks. While such methods enable emotion collection without overwhelming the student, the periodicity of the self-reports may influence the emotions being reported. Frequent prompts might interrupt the learning process and induce negative emotions like frustration, while infrequent prompts may miss short-lived emotional shifts, potentially impacting the accuracy of the detection models.

Moreover, in these approaches, the recorded emotion is treated as a single instance, without considering the duration of the emotion or the order in which emotions are experienced. This omission limits insights into the sequence of emotional states that could offer deeper understanding of emotional dynamics during learning. For instance, tracking transitions from confusion to frustration or from boredom to engagement could be valuable for designing adaptive systems that intervene at the right moments.

In (Cocea & Weibelzahl, 2010)'s study, the collection session was divided into 10-minute intervals, with each interval receiving a single label regarding students' engagement. The sessions were viewed sequentially, enabling the capture of a sequence of students' emotions. (Yang et al., 2016) collected emotions from the class forum, with each post being labeled with an indication of the confusion level.

RQ1.4 - Which theories or prior works underpin the detection of these emotions?

This RQ aims to identify the psychological theories or works on emotions that the selected articles relied on to determine which emotions to detect. Our analysis revealed that the selection of emotions for detection was typically validated based on either the emotions addressed in the protocol used to collect ground truth labels (i.e., the actual emotional states observed or reported, which serve as the reference or "true" values used to train and evaluate machine learning models) or earlier research in emotion detection.

Most selected works (Baker et al., 2014; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Paquette et al., 2016) relied on the coding scheme from the BROMP protocol. This protocol, often used in emotion detection studies, applies a 20-second observation window during which emotions are annotated. This window is used to approximate the duration of an emotion, providing a means to measure how long a particular emotional state persists. Although this method gives a rough measure of emotion duration, it may not fully capture rapid emotional shifts occurring within shorter timeframes. Nonetheless, the BROMP protocol remains a foundational tool for synchronizing log data with observed emotional states. In addition, the papers (Baker et al., 2012; Cocea & Weibelzahl, 2010; Paquette et al., 2014; Smeets et al., 2019) were based on prior research conducted by Baker and colleagues (Baker et al., 2010, 2012; Botelho et al., 2017), which also employed BROMP as a basis for their emotion coding.

Several works selected their emotions based on educational frameworks or theories of emotions. The emotion selection in the works by Hutt et al. (2019) and Jensen et al. (2019) is based on Pekrun's Control-Value Theory of emotions in education (Pekrun, 2007; Pekrun & Linnenbrink-Garcia, 2012; Pekrun et al., 2002a). Similarly, Wixon and Arroyo (2014) used the affective model developed by Kort and Picard (Kort et al., 2001). Other studies have referenced the theoretical model of emotions in deep learning proposed by Graesser (Graesser & D'Mello, 2011). For instance, the work by Yang et al. (2016) builds on previous studies by Baker and incorporates the Graesser and D'Mello model (S. D'Mello & Graesser, 2012; D'Mello et al., 2014; D. M. C. Lee et al., 2011; Lehman, D'Mello, & Graesser, 2012; Lehman & et al., 2008; Pardos et al., 2014). Similarly, Pardos et al. (2014) based their study on a range of prior works, including contributions by Baker, Graesser, and D'Mello, drawing from foundational research in emotion modeling in computer-based learning (Aleven et al., 2004; Baker, 2007; Baker et al., 2010, 2012; Cocea et al., 2009; Craig et al., 2004; D. M. C. Lee et al., 2011; Lehman, D'Mello, & Graesser, 2012; Rodrigo et al., 2009).

Some efforts, however, based their choice of emotions on prior research in emotion detection. For instance, (Bosch et al., 2015) based their emotion selection on D'Mello's meta-analysis on emotions in learning technologies (S. D'Mello, 2013) as well as on their personal observations of students during the first day of data collection. The study by (Kai et al., 2015) is grounded in D'Mello's research on the bodily expression of affect (S. D'Mello, 2011). The work of (Ocumpaugh et al., 2014) based their selection of emotions on different works for each emotion: boredom (Csikszentmihalyi, 1990; Miserandino, 1996), confusion (Craig et al., 2004; Kort et al., 2001), engagement (Csikszentmihalyi, 1990), and frustration (Kort et al., 2001; Patrick et al., 1993). Lastly, the work of (Tiam-Lee & Sumi, 2019) made no mention of their decision about the emotions considered.

RQ1.5 - Do the studies incorporate data beyond log entries to improve the detection of emotions?

Various factors can impact the transition of students' emotions during the learning process. These factors include gender, behaviors, and the duration of emotions. Gender has been found to have an influence on the range of emotions experienced by individuals (Frenzel et al., 2007; Hembree, 1988; Hyde et al., 1990), and there are observed differences in emotional appraisals between genders (Pekrun, 2016). Moreover, students' behaviors in CBLEs, such as engaging in on-task conversations or exhibiting off-task behavior, can redirect their learning trajectory (Baker et al., 2004; Baker et al., 2010). Additionally, the duration of emotions has been noted as a contributing factor to the specific emotions experienced by students and the manner in which they experience them (de Morais & Jaques, 2023; Graesser & D'Mello, 2011; Reis et al., 2018).

Therefore, this RQ aims to determine whether the authors utilized information beyond the log data to enhance the automatic recognition of student emotions. The works of (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; Cocea & Weibelzahl, 2010; Henderson et al., 2020; Hutt et al., 2019; Jensen et al., 2019; Ocumpaugh et al., 2014; Paquette et al., 2014; Smeets et al., 2019; Wixon & Arroyo, 2014; Yang et al., 2016) do not discuss the incorporation of new information in conjunction with emotions.

The papers of (DeFalco et al., 2018; Jiang et al., 2018; Kai et al., 2015; Paquette et al., 2016; Pardos et al., 2014) describe the collection of students' behavior alongside emotions. Students' behavior refers to their actions or attitudes related to the learning tasks, also known as learning-oriented behavior, which has been observed to impact learning outcomes (Perkins, 1965). For example, off-task behavior (related to not working on the content and not paying attention to the current learning task) has been found to negatively correlate with learning (S. W. Lee et al., 1999) and is associated with higher dropout rates (Finn, 1989). Conversely, on-task behavior has a positive impact on learning (Bloom, 1977; Carroll, 1963).

Only the papers of (Kai et al., 2015) and (Jiang et al., 2018) reported the development of automatic behavior recognition. Similarly, (Tiam-Lee & Sumi, 2019) collected student actions by asking students to name their state, such as reading, thinking, writing, noting, and unfocused. However, none of these papers discuss the application of this information in the development of emotion detectors.

5.2 RQ2 - What are the methodologies and technologies behind the development of sensor-free affect detectors specifically for CBLEs?

Investigating the intricacies of developing sensor-free affect detectors for CBLEs necessitates a deep dive into the methodologies and technologies that make such detection possible. RQ2 unfolds into several sub-questions designed to examine each component of the development process, from the identification of key features and algorithms to the application of specific tools and technologies. These sub-questions collectively aim to uncover the breadth of approaches used to capture, process, and analyze emotional data in educational settings without relying on physical sensors. By exploring how features are selected, which machine learning algorithms are preferred, the evaluation metrics for model performance, and the tools that facilitate these developments, this section provides a comprehensive overview of the current state and challenges in the development of sensor-free affect detection within CBLEs.

RQ2.1 - What features are predominantly used for emotion detection, and what methods are used for their capture and selection?

The modeling process of sensor-free affect detectors involves the application of machine learning techniques to predict and classify emotions. This process is reliant on the extraction of input features derived from student actions within a CBLE, typically obtained from log data. In this study, a log is defined as a recorded entry that captures pertinent data pertaining to the actions undertaken by students within the CBLE. These logs are commonly stored in the CBLE database or specific log files.

Each log can contain multiple pieces of information, with each piece representing a unique "feature" associated with the log. During the modeling process, feature engineering techniques are employed to generate additional features that encompass a wide range of characteristics and patterns found within the log data. This feature engineering step plays a crucial role in improving the accuracy of emotion detection and enables a more comprehensive understanding of student behavior (Botelho et al., 2019). Notably, there are various approaches that can be utilized in feature engineering, including the extraction of temporal patterns to analyze the evolution of emotions over time. Additionally, the incorporation of contextual information or the consideration of individual characteristics allows for the accounting of situational and personal factors that influence emotions.

This RQ aims to investigate the most frequently utilized features for detecting each emotion and to explore the methods employed for collecting these features within CBLEs. Various approaches to feature selection are employed in works on sensor-free affect detection, driven by the diverse nature of emotions and their complex manifestation in educational settings. These approaches aim to identify features that are highly informative and can effectively capture the underlying emotional states of students. By selecting the most relevant features, researchers can enhance the performance and interpretability of their emotion detection models.

The choice of features depends on multiple factors, including the specific research objectives, available data, and characteristics of the CBLE. Researchers may consider features related to student behavior, such as interaction patterns, engagement levels, or temporal dynamics. Additionally, they may explore contextual information or demographic factors to enrich the feature set further. The ultimate goal is to capture a comprehensive representation of students' emotional experiences and tailor the detection models accordingly.

Through an investigation of the most frequently utilized features for detecting each emotion and an exploration of the methods employed for their collection, valuable insights can be gained into the specific indicators and patterns associated with different emotional states. This knowledge contributes to the development of robust and accurate emotion detection models within the context of computer-based learning environments.

The works used different features depending on the type of CBLE (such as MOOCs and intelligent tutoring systems) and the subject taught (such as math and physics). For example, (Tiam-Lee & Sumi, 2019) used data from a CBLE that helps students with programming tasks. The authors made sensor-free affect detectors by considering all of the changes made to the code, such as insertions, deletions, compilations, and submissions of the assignments. (Yang et al., 2016) collected the data from a MOOC. Thus, they focused on collecting features about posts and the student's actions in the environment through their computer mouse (clickstream). In their study, (Jensen et al., 2019) and (Hutt et al., 2019) collected data from Algebra Nation, a CBLE dedicated to math education. They gathered information on 22 specific features that were independent of specific content, such as video selection or quiz questions. Data analysis involved counting the occurrences of each feature within 30-second intervals and summing these counts across 5-minute window periods, with a maximum of 10 recorded activities for each 30-second chunk. (Cocea & Weibelzahl, 2010) described the collection of data from the HTML-Tutor, which assists students in learning HTML. The authors developed emotion detectors based on features about the number of pages visited, the number of tests taken, the average time spent on pages, the average time spent on tests, the number of correctly answered tests, and the number of incorrectly answered tests.

In addition to the features that are unique to each type of CBLE and content, other works used more general features to build their sensor-free affect detectors. For example, some papers (Bosch et al., 2015; Kai et al., 2015; Paquette et al., 2016; Pardos et al., 2014; Smeets et al., 2019) described the use of features related to the number, type, and time of actions in the system, mouse clicks, keystrokes, and the aggregation of actions (e.g., the number of clicks in the last 5 seconds). Besides using general features, some works (Baker et al., 2012, 2014; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Ocumpaugh et al., 2014; Paquette et al., 2014) added temporal and skill-based features and features based on the number of errors, correct answers, and hints requested. Still, some works describe the division of the features into groups (Jiang et al., 2018), such as basic features (related to usage patterns), sequence features (related to sequential actions in the system), and threshold features (related to the amount of each feature when considering multiple students). Of all the works from this analysis, only (Wixon & Arroyo, 2014) does not present information about the features considered during the model's development.

Several different features may be used to represent a single student's action with the system. However, using too many features during training can negatively impact the machine learning models. Some reasons are issues that make the results not generalizable for unseen data, such as noisy data and high sample dimensions. Feature selection is a common way to deal with a large number of features. Its goal is to find the features that best describe the data based on the target label, in this case, the student's emotions. We could identify four feature selection approaches used by the selected works.

- **Forward Selection:** This was the most cited approach from the selected papers (Baker et al., 2012, 2014; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Paquette et al., 2014, 2016; Pardos et al., 2014). This type of selection starts with an empty set of features and adds the feature that achieves the best performance each time.

- **Tolerance Analysis:** This was the second most frequently cited method (Bosch et al., 2015; Jiang et al., 2018; Tiam-Lee & Sumi, 2019). This type of selection evaluates features' multicollinearity and eliminates highly collinear ones.

- **Correlation-based:** Used only by (Kai et al., 2015), the goal is to eliminate features with very little correlation with the target label. This approach also aims to remove features that are highly correlated with each other.

- **Backward Elimination:** Used only by (Paquette et al., 2014), this approach can be seen as the opposite of the forward selection. It begins with a list of all features and removes those that have no significant effect on performance.

In addition to reporting the approach used for feature selection, certain articles also provided details on additional aspects, including algorithms, thresholds, and validation strategies. Some papers even employed multiple strategies to enhance their affect detection models (Jiang et al., 2018; Paquette et al., 2014). However, it is worth noting that certain works (Botelho et al., 2017; Cocea & Weibelzahl, 2010; Hutt et al., 2019; Jensen et al., 2019; Ocumpaugh et al., 2014; Smeets et al., 2019; Wixon & Arroyo, 2014) did not perform or document any well-known feature selection processes, such as backward elimination. Nevertheless, some of these works utilized specific preprocessing techniques, such as summing up feature counts (Hutt et al., 2019; Jensen et al., 2019).

From the papers that reported a feature selection phase, (Baker et al., 2012, 2014; Hutt et al., 2019; Jiang et al., 2018; Kai et al., 2015; Paquette et al., 2014; Pardos et al., 2014; Smeets et al., 2019; Tiam-Lee & Sumi, 2019; Yang et al., 2016) described specific details on the selected features for each emotion model[3].

Several works provided details about the features selected for emotion detection models, with common features being identified across most detectors regardless of the specific emotion. These common features include the number and frequency of actions, such as clicks or task attempts, the time taken to perform actions, and the intervals between them. Additionally, the history of help requests and the correctness of responses, as well as the number of actions completed within different time windows, were frequently selected features.

We were also able to find some features that were chosen for a specific emotion. Models were used to detect if someone was bored by looking at the number of wrong answers in a row, how fast they happened, and the chances of getting the answer right or guessing. Confusion was linked to characteristics of task types, task difficulty, the number of consecutive incorrect actions and their speed, the number of hints requested, the number of incomplete actions, the likelihood of answering or guessing correctly, and click patterns. Delight was related to features about the number of gamification trophies, the number of completed tasks or correct answers, and the time

[3] The complete list of features related to the detection of each emotion can be found in this link: https://bit.ly/3VTOmaC.

spent on different actions. Engagement was based on the number of completed tasks or correct answers, the history of actions, and the number of hints asked for. Frustration models used features about the number of hints requested, the number of consecutive incorrect actions and their speed, and the time of inactivity.

We also looked into how researchers synchronized the action logs and the emotion labels. Only some works (Bosch et al., 2015; Cocea & Weibelzahl, 2010; Hutt et al., 2019; Jensen et al., 2019; Paquette et al., 2014, 2016; Pardos et al., 2014) have described this process. First, we have identified the grain size of the emotion labels, i.e., the assumed duration of an emotion. This duration has to do with the protocol the authors used to collect and annotate the student's emotions. The works by (Bosch et al., 2015; Paquette et al., 2014; Pardos et al., 2014) collected emotions in a 20-second window. Other works (Cocea & Weibelzahl, 2010; Hutt et al., 2019; Jensen et al., 2019) reported the emotion collection in windows of 1, 3, 5, and 10 minutes. This information is important because the synchronization between logs and emotions depends on this time. For example, suppose the work adopts a window size of 20 seconds, and a student performs 15 actions during this period. In that case, the authors must decide which approach to take for combining this information. We have identified two different approaches used by the selected papers. The first approach, used by (Paquette et al., 2016; Pardos et al., 2014), replicates the emotion in these 15 action logs. Thus, each log inside the 20-second window will have the same synchronized emotion. The second approach, used by (Bosch et al., 2015; Pardos et al., 2014), aggregates the logs through some computation. For instance, the number of clicks, the average time spent on each action, and the number of hints asked for. So, the 15 logs will be combined into a single log with their information and a label for the emotion.

Another investigation we have performed is about the missing values. In a log, each feature represents different information. Some actions may not have enough information to fill in all the features. (Bosch et al., 2015; Kai et al., 2015) described the presence of missing values and the application of three different approaches to dealing with them. The first approach is zero imputation, used by (Bosch et al., 2015; Kai et al., 2015), in which every single missing value is set to zero. The second approach is average imputation, used by (Kai et al., 2015), which takes the non-missing values of the same feature, computes their average, and sets the missing values to the computed average. The third approach, used by (Kai et al., 2015), is single imputation, in which the authors have built a tree-based decision model (M5) to predict the best value for the missing values of each feature.

RQ2.2 - Which machine learning algorithms are primarily employed in sensor-free emotion detection?

We have also investigated the machine learning algorithms employed in the selected works, analyzing their development and analysis processes. While the number of tested algorithms varies, all the chosen works explicitly cite the algorithms they used.

We could observe that the common approach to developing sensor-free affect detectors is by starting with a set of algorithms, then selecting one model according to some goal. Even though some works did not provide any explanation for the algorithms' selection (Bosch et al., 2015; Hutt et al., 2019; Tiam-Lee & Sumi, 2019; Wixon & Arroyo, 2014; Yang et al., 2016), most papers (Baker et al., 2012, 2014; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Pardos et al., 2014) reported that they chose a specific set of common or standard classification algorithms that had previously been shown to be successful in building emotion detectors. Another reason is that the chosen algorithms can show different patterns in the data, but they are fairly conservative and less likely to overfit. (Botelho et al., 2017) described the selection of three common recurrent network variants. Furthermore, some works selected their algorithms based on previous research, such as (DeFalco et al., 2018; Paquette et al., 2014, 2016; Smeets et al., 2019) citing (Baker et al., 2012), (DeFalco et al., 2018; Paquette et al., 2014, 2016) citing (Pardos et al., 2014), (Henderson et al., 2020) citing (DeFalco et al., 2018), (Jensen et al., 2019) citing (Hutt et al., 2019), and (Cocea & Weibelzahl, 2010) citing (Mitchell & Mitchell, 1997; Witten et al., 2005).

We also investigated the strategy they used to select the models. We have identified three distinct strategies. The first and most commonly used strategy (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; Henderson et al., 2020; Jiang et al., 2018; Kai et al., 2015; Paquette et al., 2014, 2016; Smeets et al., 2019; Tiam-Lee & Sumi, 2019) is to select the model that achieves the best performance for each emotion. In this scenario, if the paper considers four different emotions, the authors choose the best model for each emotion. The second strategy selects the models according to a sample of the data. For instance, (Ocumpaugh et al., 2014) selected the best model according to the data from different populations. Finally, the third strategy selects the model based on its successful application in previously developed work (Cocea & Weibelzahl, 2010; Jensen et al., 2019). In this case, the authors do not test any model. Instead, they choose the models according to the reported performance from previous work. This strategy differs from the first one in that it relies on the results of a model applied in a previous publication, rather than executing the model for the current work. It is worth noting that if the previously published work is by the same authors, the distinction between the first and third strategies may not be significant. However, we believe it is important to differentiate these strategies for the sake of clarity and thoroughness in our analysis.

The majority of papers developed single-class (or binary) classification models (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2014, 2016; Pardos et al., 2014; Smeets et al., 2019; Tiam-Lee & Sumi, 2019; Yang et al., 2016). In this case, the authors have trained the classification models to learn the presence or absence of a single emotion, for example, to identify whether the student is confused or not. Some works have followed a multi-classification approach, in which the models are trained to differentiate between multiple classes (Botelho et al., 2017; Cocea & Weibelzahl, 2010). The model outputs whether the student is engaged, confused, neutral, and so on. (Jensen et al., 2019; Wixon & Arroyo, 2014) trained the models to infer the valence of the student's emotions, i.e., report whether the student is experiencing a positive or negative emotion. Finally, (Hutt et al., 2019; Wixon & Arroyo, 2014) inferred the intensity of each emotion or the level of presence or absence of an emotion, a regression problem.

During our investigation of the selected works, we examined the strategies employed for hyperparameter tuning and algorithm development in order to enhance the performance of sensor-free affect detectors. Hyperparameter optimization is a key focus in this process, aiming to identify the optimal parameter settings for a learning algorithm and thereby improving the model's overall performance. The incorporation of hyperparameter tuning and algorithm development within a systematic literature review on sensor-free affect detection is crucial, as it ensures that the advancements achieved through these techniques are thoroughly documented and analyzed. By identifying the specific approaches employed in the reviewed works, researchers and practitioners can gain valuable insights into the most effective strategies for optimizing affect detection models and driving advancements in the field.

Among the selected papers, (Botelho et al., 2017; Hutt et al., 2019; Jiang et al., 2018; Smeets et al., 2019) adopted hyperparameter optimization or fine-tuning techniques to optimize their models. Both (Smeets et al., 2019) and (Hutt et al., 2019) employed the grid search strategy with cross-validation. This approach involves systematically exploring a predefined set of parameters for each algorithm, training and evaluating multiple models with different parameter combinations. Ultimately, the configuration that yields the best performance is selected as the final model.

In addition to grid search, (Botelho et al., 2017; Hutt et al., 2019; Jiang et al., 2018) utilized various techniques to refine the parameters of their neural network models. These techniques included genetic algorithms, dropout regularization, and adjustments to the network topology. By employing these strategies, the authors achieved further optimization of the model's parameters, resulting in improved accuracy in detecting and classifying emotions.

RQ2.3 - What evaluation metrics are utilized to assess the quality of the models?

To evaluate and report on the performance of sensor-free affect detectors, researchers utilize various evaluation metrics depending on the type of problem, such as classification or regression. In general, we have observed that papers adopting the classification approach tend to employ Cohen's Kappa and/or AUC as their primary evaluation metrics (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2014, 2016; Pardos et al., 2014; Smeets et al., 2019; Wixon & Arroyo, 2014; Yang et al., 2016). These metrics are preferred as they yield more reliable results, particularly when dealing with unevenly distributed data across different classes. On the other hand, papers focusing on the intensity of emotions captured by instruments, which entails a regression problem, utilize correlation metrics to compare the inferred intensity with the intensity captured by the instruments (Hutt et al., 2019; Jensen et al., 2019), often measured using Likert scales.

RQ2.4 - What approaches are used for training and developing the machine learning models?

We found that most works used the **K-fold cross-validation strategy** to train and validate the models. Cross-validation is a resampling method that splits the data into training and testing subsets, and it is essential for improving the generalizability of the models. This strategy divides the entire database into k mutually exclusive training and testing sets, with detectors trained and tested in k rounds. In each round, data from $k-1$ groups is used to train the detectors, and data from the remaining group is used to test the detectors. Cross-validation is important because it allows for more reliable estimates of model performance by using different data subsets in each round, reducing the risk of overfitting to a specific dataset. Based on selected works, the value for k varies between 10 (DeFalco et al., 2018; Hutt et al., 2019; Jensen et al., 2019; Jiang et al., 2018; Kai et al., 2015; Paquette et al., 2016; Tiam-Lee & Sumi, 2019; Yang et al., 2016), 6 (Baker et al., 2012), 5 (Baker et al., 2014; Botelho et al., 2017; Ocumpaugh et al., 2014; Pardos et al., 2014; Smeets et al., 2019; Wixon & Arroyo, 2014), and 4 (Henderson et al., 2020).

A different approach was applied in (Bosch et al., 2015; Paquette et al., 2014), where the authors used **leave-one-out cross-validation**. This method is similar to k-fold cross-validation, but instead of dividing the data into *k* subsets, each round leaves one student out as the testing set, with the rest used for training. Leave-one-out cross-validation is useful when working with smaller datasets, as it allows each data point to be used for validation, though it may be computationally expensive.

Only Cocea and Weibelzahl (2010) did not report the validation process.

Some papers provided further details about the development of their machine learning pipeline, pointing out that pre-processing steps such as data normalization, class imbalance resampling, feature selection, and model training were performed within each cross-validation fold. This approach is important to avoid data leakage between the training and testing sets, which could otherwise lead to overly optimistic performance estimates. Another strategy used by selected works is the use of a validation set, which involves splitting the data into training, validation, and testing (*hold-out*) subsets. The training and validation datasets, also called the development set, are used during the model's training. With cross-validation, these two datasets are automatically split according to the *k* value. On the other hand, the hold-out dataset is never used during model development, which prevents data leakage during the final performance analysis.

The validation level was not explicitly reported by Cocea and Weibelzahl (2010), Tiam-Lee and Sumi (2019), and Yang et al. (2016). However, all other studies employed **student-level validation**, a method in which data from each student is either included in the training set or the testing set, but never in both. This approach ensures that the model is tested on completely new students, providing stronger evidence that it can generalize well to unseen individuals.

RQ2.5 - Which tools and technologies are predominantly used in the development of these models?

RapidMiner (Baker et al., 2012; DeFalco et al., 2018; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2016; Wixon & Arroyo, 2014) and WEKA (Bosch et al., 2015; Cocea & Weibelzahl, 2010; Tiam-Lee & Sumi, 2019) are the most commonly used tools. These tools offer a range of machine learning functionalities for data science and data mining tasks.

In addition to RapidMiner and WEKA, selected works have utilized various Python libraries such as Theano (Botelho et al., 2017), Lasagne (Botelho et al., 2017), scikit-learn (Hutt et al., 2019; Jensen et al., 2019), TensorFlow (Hutt et al., 2019), and Keras (Hutt et al., 2019).

It is worth mentioning that some studies (Henderson et al., 2020; Paquette et al., 2014; Pardos et al., 2014; Smeets et al., 2019; Yang et al., 2016) did not explicitly describe the tools or technologies used in their research.

RQ2.6 - What is the range of emotion labels, action logs, and features considered in these studies?

By analyzing the selected works, there is no standard quantity of labels and features, and it varies for each work.

First, we investigated the number of logs collected and used by the selected works. Only two papers reported the number of logs examined: (Jiang et al., 2018) (146,000) and (Cocea & Weibelzahl, 2010) (1,000). On the other hand, only two papers did not present the number of features used to develop their detection models: (Tiam-Lee & Sumi, 2019; Wixon & Arroyo, 2014). All other selected works have reported the number of features considered. Again, this number is different for each work, ranging from 5 to 7 for (Baker et al., 2014), from 20 to 40 for (Cocea & Weibelzahl, 2010; DeFalco et al., 2018; Henderson et al., 2020; Hutt et al., 2019; Jensen et al., 2019; Paquette et al., 2016), from 40 to 60 for (Smeets et al., 2019; Yang et al., 2016), from 60 to 80 for (Kai et al., 2015; Ocumpaugh et al., 2014), from 110 to 130 for (Bosch et al., 2015; Paquette et al., 2014), 172 for (Pardos et al., 2014), 204 for (Botelho et al., 2017), 232 for (Baker et al., 2012), and 249 for (Jiang et al., 2018).

We also examined the distribution of the number of emotion labels used in the studies. This range varied from less than 1,000 (Baker et al., 2012) to between 1,000 and 2,000 (Bosch et al., 2015), 2,000 and 3,000 (Baker et al., 2014; Kai et al., 2015), 3,000 and 4,000 (DeFalco et al., 2018; Henderson et al., 2020; Paquette et al., 2016; Pardos et al., 2014), 4,000 and 5,000 (Paquette et al., 2014), 5,000 and 6,000 (Jiang et al., 2018; Yang et al., 2016), and 7,000 and 8,000 (Botelho et al., 2017). Notably, one study reported an unusually high number of over 133,000 emotion labels (Hutt et al., 2019), which was identified as an outlier and subsequently removed from our analysis.

Excluding the outlier, the average number of collected emotion labels across the studies was approximately 3.5,000. However, it should be noted that some studies utilized only a subset of the collected labels, with the number ranging from less than 1,000 (DeFalco et al., 2018; Henderson et al., 2020; Paquette et al., 2016) to between 1,000 and 2,000 (Bosch et al., 2015; Kai et al., 2015; Paquette et al., 2014). On average, the number of emotion labels used in the studies was approximately 2.3,000. Notably, the papers by (Cocea & Weibelzahl, 2010; Jensen et al., 2019; Ocumpaugh et al., 2014; Smeets et al., 2019; Tiam-Lee & Sumi, 2019; Wixon & Arroyo, 2014) did not provide explicit information regarding the number of emotion labels used in their respective studies.

In addition to the total number of emotion labels, we looked into how the works deal with unequal classes. This is an important subject to consider because, in classification, different numbers of instances representing each class, in this case, the student's emotions, can impose a load-balancing problem[4], leading to wrong results or interpretations of the results. Selected works used two resampling strategies to deal with imbalanced classes. The first strategy is known as "oversampling", applied by (Baker et al., 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Jiang et al., 2018; Kai et al., 2015; Paquette et al., 2016), which creates samples of the minor representative emotions. They used two approaches to create the new samples. The first one is to clone the minority samples, applied by (Baker et al., 2014; Botelho et al., 2017; DeFalco et al., 2018; Jiang et al., 2018; Kai et al., 2015), and the second one is through data augmentation, using the SMOTE technique, applied by (Bosch et al., 2015; Henderson et al., 2020). Another strategy to deal with imbalanced data is downsampling or undersampling, applied by (Bosch et al., 2015; Botelho et al., 2017; Smeets et al., 2019), which consists of reducing the sample from the majority class. As the authors of SMOTE (Chawla et al., 2002) suggested, the best way to use SMOTE would be to randomly undersample the class that makes up the majority.

[4] More details about class imbalance problem in classification tasks can be found in (Ali et al., 2013).

This approach was applied in one selected work (Bosch et al., 2015). (Baker et al., 2012; Paquette et al., 2014; Pardos et al., 2014) reported some resampling, but did not describe which they used. Besides, (Cocea & Weibelzahl, 2010; Hutt et al., 2019; Jensen et al., 2019; Ocumpaugh et al., 2014; Tiam-Lee & Sumi, 2019; Wixon & Arroyo, 2014; Yang et al., 2016) did not present any information about resampling or data imbalance.

RQ2.7 - What are the highest-performing results and corresponding algorithms for each detected emotion according to the studies?

This RQ aims to present the best models reported in each selected work. To answer this question, we analyzed each work, searching for the best result in detecting each emotion. Some studies reported different results when considering different types of data altogether (multi-modal combining interaction data with video, text, sensors, etc.). However, we considered only the results based on interaction data. Because we have so many emotions (25 emotions according to RQ1.1) and so many performance metrics (11 metrics according to RQ2.4), we selected the most frequent emotions, reporting their results with the most commonly used metrics. So, for emotions, we looked at boredom, confusion, engagement, and frustration, with more than 16 occurrences each (see Section 5.1). Figure 4 presents the AUC values for the detection of boredom, confusion, engagement, and frustration, and the algorithms that achieved each value. The color of each point identifies the results of one of the selected works in a given metric. Papers that did not report the AUC metrics for one of the four emotions were not considered.

It is important to notice that each work uses different information to figure out what emotions someone is feeling and gets its data from different CBLEs, contexts, and contents. Therefore, Figure 4 aims to provide an overview of the results presented in the sensor-free affect detection field. In addition, the list of algorithms is not extensive because it only shows the algorithms that achieved the best results. RQ2.2 shows the complete list of algorithms (see Section 5.2).

5.3 RQ3 - What are the primary contexts in which sensor-free emotion detectors are employed within CBLEs?

Understanding the diverse contexts in which sensor-free emotion detectors are implemented within CBLEs is essential for appreciating their impact and scalability across various educational settings. In this case, “primary contexts” refer to the main environments or educational platforms, such as intelligent tutoring systems, online learning environments, or massive open online courses (MOOCs), where sensor-free emotion detection is applied. These detectors rely on students’ interactions with the system, such as mouse clicks, keystrokes, and task performance, to infer emotional states. Each context has distinct characteristics that may influence how emotion detection is implemented. To thoroughly explore these contexts, RQ3 is divided into five sub-questions, each aimed at illuminating a different aspect of the application environment. These sub-questions collectively seek to delineate the characteristics of the samples involved, the types of CBLEs in use, and the learning domains these systems cover.

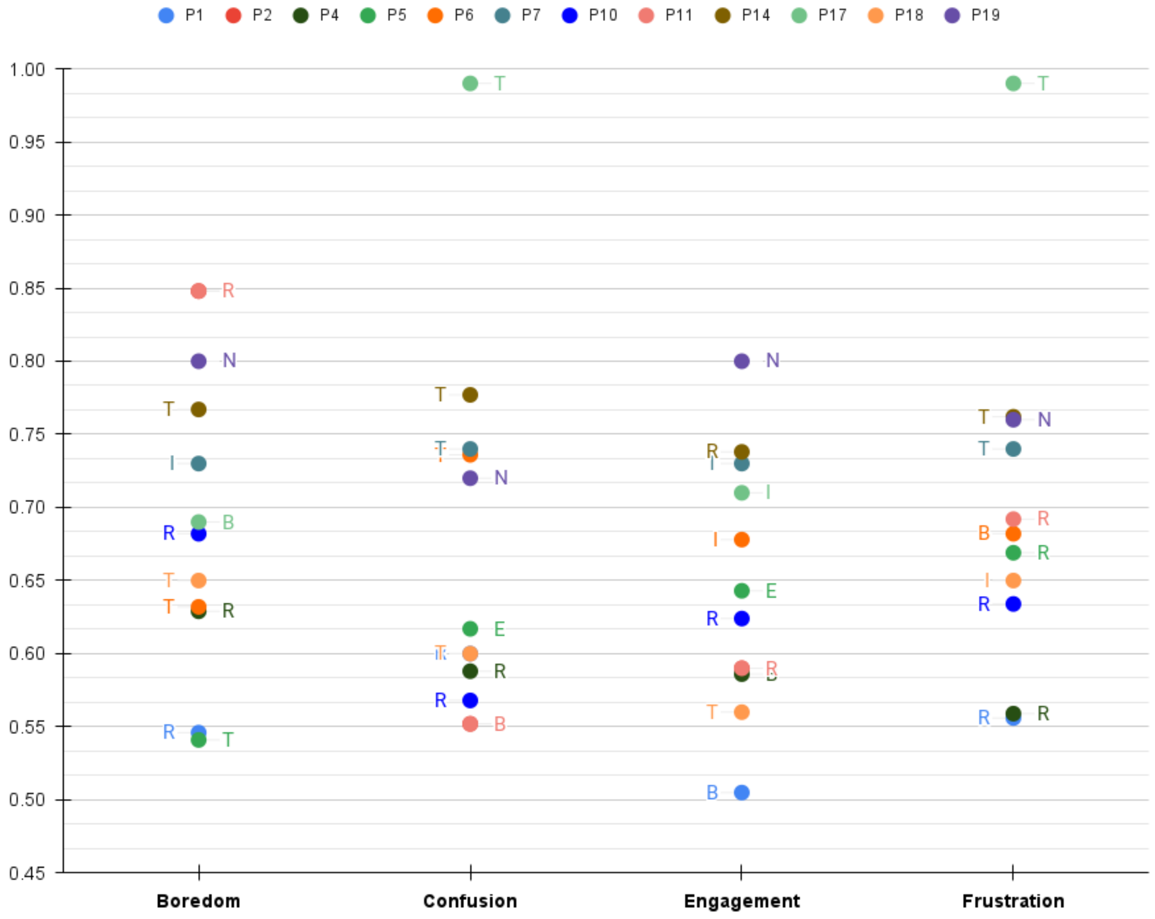

Figure 4: Scatter plot showing the best results for each study on sensor-free affect detectors, measured in AUC. Each point on the graph represents the optimal result found in a study. The labels next to the points represent the type of algorithm that achieved the highest result, which are: **R** = Regression , **B** = Bayesian , **T** = Decision Tree , **E** = Ensemble-based , **I** = Instance-based , **N** = Neural Networks..

RQ3.1 - What sample characteristics are reported in the studies?

This RQ identifies the sample in which the data is being collected. Most research has been conducted on regular K–12 education. The works have collected data from 6th (Jiang et al., 2018; Smeets et al., 2019), 8th (Paquette et al., 2014; Pardos et al., 2014), 8th to 9th (Bosch et al., 2015; Kai et al., 2015), 7th to 10th (Wixon & Arroyo, 2014), and 6th to 12th grades (Jensen et al., 2019). Works also collected data from students in the military academy (DeFalco et al., 2018; Henderson et al., 2020; Paquette et al., 2016) and 1st year programming classes (Tiam-Lee & Sumi, 2019). The papers by (Baker et al., 2012, 2014; Botelho et al., 2017; Cocea & Weibelzahl, 2010; Hutt et al., 2019; Ocumpaugh et al., 2014; Yang et al., 2016) did not provide information about the students' sample.

We also looked into the sample size or the number of students from whom the action logs were taken. We found out that papers collected data from the following number of students: less than 50 (Cocea & Weibelzahl, 2010), between 73 and 93 (Baker et al., 2012; Jiang et al., 2018; Tiam-Lee & Sumi, 2019), between 119 and 153 (Baker et al., 2014; Bosch et al., 2015; DeFalco et al., 2018; Henderson et al., 2020; Kai et al., 2015; Paquette et al., 2016), between 229 and 326 (Paquette et al., 2014; Pardos et al., 2014; Wixon & Arroyo, 2014), 646 (Botelho et al., 2017), 4281 (Yang et al., 2016), and 69174 students (Hutt et al., 2019; Jensen et al., 2019).

Only three papers reported the students' age, in which (Smeets et al., 2019) collected data from students between 4 and 12 years old and (DeFalco et al., 2018; Henderson et al., 2020) from students between 18 and 22 years old.

Most of the selected works collected data from students using the CBLE in the school's lab (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Hutt et al., 2019; Jensen et al., 2019; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2014, 2016; Pardos et al., 2014; Tiam-Lee & Sumi, 2019; Wixon & Arroyo, 2014; Yang et al., 2016). (Smeets et al., 2019; Yang et al., 2016) reported collecting action logs from students using the CBLE at home, whereas (Cocea & Weibelzahl, 2010) did not specify where the students were while using the CBLE.

Of the works that collected data from schools, only (Bosch et al., 2015; Jiang et al., 2018; Kai et al., 2015) described students as attending a public school. (Smeets et al., 2019) also described that students paid for a subscription to use the CBLE. Other selected works did not contain this information.

Some studies have reported additional information about the sample, such as genders (Bosch et al., 2015; DeFalco et al., 2018; Henderson et al., 2020; Kai et al., 2015; Paquette et al., 2016), ethnicities (Pardos et al., 2014), schools (Hutt et al., 2019; Jensen et al., 2019; Paquette et al., 2014), and countries (Tiam-Lee & Sumi, 2019).

We also analyzed the period over which the selected works collected student log data. The results show that it varies from a few days (1 to 7 days) (Baker et al., 2012, 2014; Bosch et al., 2015; Cocea & Weibelzahl, 2010; Jiang et al., 2018; Kai et al., 2015; Paquette et al., 2016; Wixon & Arroyo, 2014), weeks (Yang et al., 2016), and a whole school year (Botelho et al., 2017; Hutt et al., 2019; Jensen et al., 2019; Ocumpaugh et al., 2014; Pardos et al., 2014). Regarding the duration of the sessions, they vary between 30 and 45 minutes (Jiang et al., 2018; Tiam-Lee & Sumi, 2019), 55 minutes (Bosch et al., 2015; Kai et al., 2015), and one to two hours (DeFalco et al., 2018; Henderson et al., 2020; Paquette et al., 2016; Pardos et al., 2014). (Paquette et al., 2014; Smeets et al., 2019) did not provide this information.

We also identified where the sample came from. Most of the sensor-free affect detection research collected data from students in the United States (Baker et al., 2012, 2014; Bosch et al., 2015; Botelho et al., 2017; DeFalco et al., 2018; Henderson et al., 2020; Hutt et al., 2019; Jensen et al., 2019; Jiang et al., 2018; Kai et al., 2015; Ocumpaugh et al., 2014; Paquette et al., 2014, 2016; Pardos et al., 2014; Wixon & Arroyo, 2014; Yang et al., 2016). A few studies were done in other countries, like the Netherlands (Smeets et al., 2019), Japan (Tiam-Lee & Sumi, 2019), the Philippines (Tiam-Lee & Sumi, 2019), and Germany (Cocea & Weibelzahl, 2010).

RQ3.2 - What specific types of CBLEs, such as ITS, MOOCs, among others, are being utilized?

The selected works prominently feature CBLEs that can be categorized into several types, including Intelligent Tutoring Systems (ITS), game-based environments, web-based platforms, Massive Open Online Courses (MOOCs), coding IDEs, and virtual worlds.

In the ITS category, we encountered a diverse range of systems, such as Inq-ITS (Paquette et al., 2014), HTML-Tutor (Cocea & Weibelzahl, 2010), ASSISTments (Botelho et al., 2017; Ocumpaugh et al., 2014; Pardos et al., 2014), Cognitive Tutor Algebra I (Baker et al., 2012), Betty's Brain (Jiang et al., 2018), and Wayang Outpost (Wixon & Arroyo, 2014).

Transitioning our attention to game-based environments, we identified various platforms like Physics Playground (Bosch et al., 2015; Kai et al., 2015), vMedic - TC3Sim (DeFalco et al., 2018; Henderson et al., 2020; Paquette et al., 2016), and Squla (Smeets et al., 2019).

Our research also revealed distinct CBLEs that were the focus of individual studies. These include the virtual world EcoMUVE (Baker et al., 2014), a customized Integrated Development Environment for coding (Tiam-Lee & Sumi, 2019), and an investigation conducted on the MOOC platform, Coursera (Yang et al., 2016). Furthermore, we encountered two studies that utilized data from Algebra Nation (Hutt et al., 2019; Jensen et al., 2019), a web-based platform designed to support algebra learning, which integrates instructional videos, practice exercises, and group discussions.

RQ3.3 - In which specific learning domains or content areas are these CBLEs being applied?

This RQ investigates the learning domain in which the selected works apply or develop their sensor-free affect detectors. We found out that math is the most studied subject, with nine works (Baker et al., 2012; Botelho et al., 2017; Hutt et al., 2019; Jensen et al., 2019; Ocumpaugh et al., 2014; Pardos et al., 2014; Smeets et al., 2019; Wixon & Arroyo, 2014; Yang et al., 2016). The second and third most studied subjects are medicine (DeFalco et al., 2018; Henderson et al., 2020; Paquette et al., 2016) and science (Baker et al., 2014; Jiang et al., 2018; Paquette et al., 2014), with three works each. The fourth and fifth most studied subjects are programming (Cocea & Weibelzahl, 2010; Tiam-Lee & Sumi, 2019) and physics (Bosch et al., 2015; Kai et al., 2015), with two works each. The final subject on the list is economics, which is taught together with mathematics in work described in the paper (Yang et al., 2016).

5.4 RQ4 - How do sensor-free emotion detectors perform in terms of generalization and practical application within CBLEs, and what does the current research landscape reveal about trends, key findings, and future directions?

RQ4 is centered on evaluating the generalization capabilities of sensor-free emotion detectors and their readiness for practical deployment in computational-based learning environments. Additionally, this question seeks to illuminate the current trends within the research community, identifying key discoveries and potential paths for future investigations. To thoroughly explore these areas, RQ4 has been divided into three sub-questions. These subdivisions are designed to assess the practical applicability and generalization performance of the detectors, examine prevailing research trends and future directions in the field, and provide an overview of the research landscape, including dominant publication venues, prolific authors, and the evolution of research over time.

RQ4.1 - What is the generalization performance of the developed detectors and how could they be applied in production?

This RQ verifies what the selected works described about the generalization performance of their models for different learners' profiles. We also examined whether the selected works applied their models in production to detect students' emotions in real-time. We found that only seven works, or 37%, mentioned how the models could be used to detect emotions for other learners' profiles. Two of these studies say that the results cannot be generalized because the sample size was too small (Tiam-Lee & Sumi, 2019) or there were too few courses (Yang et al., 2016).

Other works presented to what extent their results are generalizable. (Paquette et al., 2016) described that their posture-based detectors might achieve a better generalization performance because they consider the learner behavior outside the software, making the detectors system-independent. (Ocumpaugh et al., 2014) collected data from different populations. They showed that the detection models could only be used for different populations if they were included in the sample. (Jensen et al., 2019) trained models over a different number of students and found that models trained with fewer than 1500 students did not generate stable scores or predictions. (Cocea & Weibelzahl, 2010) mentioned that the proposed approach is generalized to systems other than e-learning. (Hutt et al., 2019) used trained models on data from different content, demonstrating that models trained for one content could be used on data for another.

Only one of the chosen works (DeFalco et al., 2018) talked about how the developed sensor-free detection models could be used in production. Their models were integrated into a CBLE and used by the students in an experiment. This experiment aimed to respond to the students' frustration through feedback messages. The authors tried out three different kinds of motivational feedback to determine which helped students learn the most.

RQ4.2 - What are the trends and future directions in sensor-free affect detection?

This research question (RQ) seeks to identify the current trends and potential future directions in sensor-free affect detection, as outlined in the selected works. “Trends” and “future works” are two distinct aspects in scientific research. “Trends” refer to the current or emerging directions within a field, reflecting ongoing developments and noteworthy patterns. These can include novel findings, methodologies, technologies, or approaches that have gained prominence. On the other hand, “future works” refer to areas that require further investigation or exploration. They represent the gaps, unresolved issues, or unanswered questions within the existing body of knowledge. Future works are proposed as potential research endeavors based on the limitations of the current study, preliminary findings, or opportunities for expanding upon existing knowledge.

To identify the trends, we first analyzed the authors' suggestions when reporting their results. We then synthesized the main trends into six categories, listed in order of their frequency:

- **Generalization:** This trend focuses on collecting or testing the detection models across different populations, CBLEs, and contexts (Botelho et al., 2017; Ocumpaugh et al., 2014; Paquette et al., 2016; Wixon & Arroyo, 2014; Yang et al., 2016).
- **Multimodal:** This trend involves using multiple data sources (beyond just interaction logs) to enhance the accuracy of detecting students' emotions (Bosch et al., 2015; Henderson et

al., 2020; Kai et al., 2015; Paquette et al., 2016; Tiam-Lee & Sumi, 2019).

- **Intervention:** This trend pertains to works aiming to provide meaningful and effective interventions based on the detected emotions, such as emotional feedback, content adaptation, teacher assistance, etc. (Baker et al., 2014; DeFalco et al., 2018; Pardos et al., 2014; Smeets et al., 2019).
- **Knowledge Discovery:** This trend involves the discovery of new knowledge, constructs, and understanding of students' emotions based on the developed models (Baker et al., 2012, 2014; Hutt et al., 2019; Jiang et al., 2018).
- **Real-time Detection:** This trend is derived from works that describe the integration of the models into CBLEs for real-time detection and decision-making during student learning (Hutt et al., 2019; Pardos et al., 2014; Yang et al., 2016).
- **Performance Improvement:** This trend identifies works aiming to enhance the performance of detection models using different technologies or techniques (Jensen et al., 2019; Paquette et al., 2014).

We also identified future works reported by the selected papers to provide an overview of potential directions in the field. The following future research suggestions were identified: Utilize multi-modal data sources (Bosch et al., 2015; Henderson et al., 2020; Kai et al., 2015); Implement real-time interventions and adaptive interfaces (redesign the CBLE based on detected student emotion) (Baker et al., 2014; Bosch et al., 2015; Paquette et al., 2014; Tiam-Lee & Sumi, 2019; Yang et al., 2016); Enhance data mining performance (DeFalco et al., 2018; Jensen et al., 2019; Paquette et al., 2016), by applying Ensemble approaches (Hutt et al., 2019), Deep learning (Botelho et al., 2017), Feature construction/selection (Baker et al., 2012; DeFalco et al., 2018; Hutt et al., 2019; Jensen et al., 2019; Smeets et al., 2019; Tiam-Lee & Sumi, 2019),Algorithm selection (Smeets et al., 2019), Aggregation methods (Baker et al., 2012), and Improved sampling strategies (Henderson et al., 2020); Understand how emotions influence learning and vice-versa (Baker et al., 2012, 2014; Hutt et al., 2019; Ocumpaugh et al., 2014; Pardos et al., 2014); Utilize student's affect and behavior data for professor recommendations and interventions (Pardos et al., 2014); Validate on broader populations/generalizations, contexts, and CBLEs (Baker et al., 2012; Botelho et al., 2017; Hutt et al., 2019; Jensen et al., 2019; Jiang et al., 2018; Ocumpaugh et al., 2014; Wixon & Arroyo, 2014); and Collect more samples of underrepresented emotions, such as confusion and frustration, in the classroom (Botelho et al., 2017; Yang et al., 2016).

RQ4.3 - What is the current status/overview of the research area (publication avenues, year, authors)?

This research question aims to provide an overview of the research area. We examined the journals and conference proceedings in which the selected papers were published, as well as the number of publications per year and country. Figure 5 illustrates these findings.

The conferences and journals frequently used by authors in this area include: International Conference on Educational Data Mining (EDM), International Conference on Artificial Intelligence in Education (AIEd), International Conference on Intelligent Tutoring Systems (ITS), International Conference on User Modeling, Adaptation, and Personalization (UMAP), International

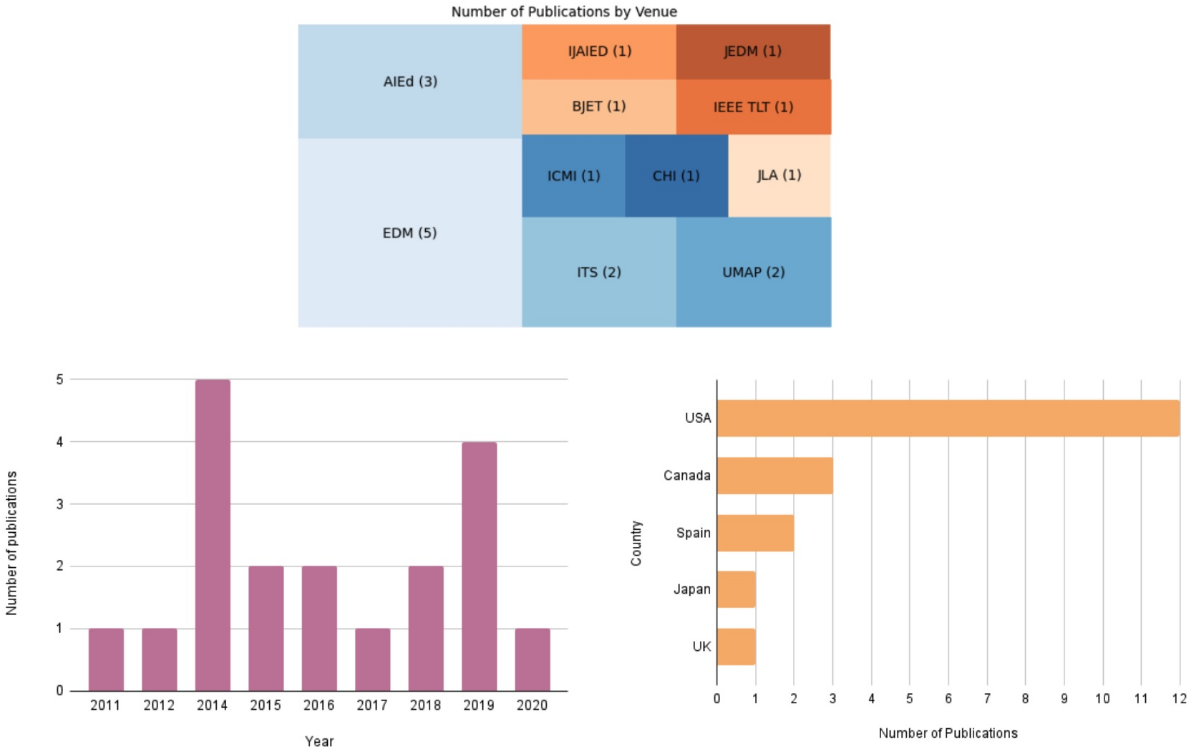

Figure 5: Overview of publication venues, countries, and years.

Conference on Multimodal Interaction (ICMI), Conference on Human Factors in Computing Systems (CHI), Journal of Learning Analytics (JLA), British Journal of Educational Technology (BJET), International Journal of Artificial Intelligence in Education (IJAIED), IEEE Transactions on Learning Technologies (IEEE TLT), and Journal of Educational Data Mining (JEDM). The frequency in each publication vehicle is illustrated Figure 5.

We compiled the names of all authors from the selected works, totaling 55 researchers. The leading authors in this field are Ryan Baker (12 publications), Luc Paquette (6 publications), Jaclyn Ocumpaugh (5 publications), Sidney K. D'Mello (4 publications), and Sujith M. Gowda (4 publications). The remaining researchers have three or fewer publications on this subject.

6 Main Findings

This section synthesizes key findings from our Systematic Literature Review (SLR) on sensor-free affect detection. Boredom, confusion, frustration, and engagement are the primary emotions detected in 84% of the 19 selected studies. However, further research into other emotions contingent on the learning environment type is encouraged. For example, (de Morais & Jaques, 2023) reported frequent instances of surprise in their step-based tutoring system.

The reviewed studies adopted various methods to identify learners' emotions, including human observer annotations, student self-reports, crowdsourcing, and log file annotation. The Behavioral Observation of Students in Schools (BROMP) protocol is often employed for encoding students' emotions in real-time, mainly due to its creators' high citation rate. However, it precludes the acquisition of sequential emotional data due to the round-robin annotation approach. Self-reporting is the second most prevalent method, followed by crowdsourcing. Some studies used log file annotation, which relies solely on log data to annotate emotions.

When developing sensor-free emotion classifiers, it's crucial to consider the "grain size," or the time duration during which emotions are annotated, and actions are captured. As discussed in Section 5.1, this duration varies significantly across studies, affecting classifier performance and being constrained by the annotation method.

The Control-Value Theory (CVT) is the most frequently referenced psychological theory for defining emotions in the selected works. CVT centers on learning-oriented emotions and suggests emotions originate from individuals' situational assessments in relation to their objectives (as noted in Section 5.1).

In developing sensor-free detection models, the selected studies utilized advanced data mining techniques, including feature selection approaches, resampling strategies, and strategies for hyperparameter tuning and evaluation such as grid search and k-fold cross-validation. These efforts enhance model performance and generability. The typical approach involves starting with a set of algorithms, then selecting a model based on a specific goal, usually superior performance according to one or more metrics. The most commonly used metrics are Cohen's Kappa and AUC.

The studies employed diverse algorithms to construct detectors. Despite the difficulties in direct comparisons due to the varying CBLEs, features, and emotion annotation methods, neural networks and decision tree algorithms generally yield the best results (as depicted in Figure 4).

Regarding features, prominent features for detecting specific emotions were identified during feature engineering. Boredom was detected by considering features such as the number of consecutive incorrect answers and their speed. Confusion was linked to task types, task difficulty, the number of hints requested, and click patterns, among others. Delight was related to the number of gamification trophies, completed tasks, correct answers, and the time spent on various actions. Engagement and frustration were deduced from features including the number of completed tasks or correct answers, action history, the number of hints requested, and inactivity periods.

Most sensor-free emotion detection research involves K–12 students in the U.S., utilizing CBLEs in school labs. Most studies have data from several hundred students, with only two studies collecting data from over 1,000 students. (Jensen et al., 2019) noted models trained with fewer than 1500 students lacked stable scores or predictions.

A significant proportion of the research was conducted using data from intelligent tutoring systems, reflecting the research community's strong background in artificial intelligence in education. The publication venues predominantly belong to the same field.

On model generalization, (Ocumpaugh et al., 2014) noted that models are only applicable to different populations if trained with samples from the targeted population. As for generalizing models to other environments, existing results are preliminary.

7 Future Research Directions

This systematic literature review has highlighted the significant progress made in the field of sensor-free affect detection in CBLEs. However, it also underscores several areas that warrant further exploration. These areas can be broadly categorized into three types: improving model performance, enhancing model development practices and methods, and integrating models into CBLEs.

1. **Improving Model Performance**: The current models have shown promise in detecting students' emotions in CBLEs. They have utilized a range of machine learning techniques and have been trained on various types of student interaction data. However, there is room for improvement in their performance. Future research directions in this area could include:

    - **Enhancing the performance of the models**: While advanced neural network algorithms have been explored in some studies, such as (Botelho et al., 2017, 2019), these works utilized emotion labels collected via the BROMP protocol. This protocol annotates emotions in a round-robin fashion, moving to the next student following an annotation. Consequently, these studies do not capture the sequence of emotions experienced by individual students. Given that research indicates the existence of an emotional dynamic among students (S. D'Mello & Graesser, 2012), incorporating this dynamic into the models could potentially enhance their emotion detection capabilities.
    - **Collect more samples of underrepresented emotions**: Emotions like confusion and frustration are often underrepresented in datasets because many CBLEs are gamified and involve small, incremental steps in tasks. These smaller steps typically reduce the occurrence of intense negative emotions, leading to fewer instances of emotions like confusion or frustration being recorded. As a result, the datasets used for model training are often imbalanced, with an overrepresentation of emotions such as engagement or delight. Collecting more samples of underrepresented emotions is crucial for improving the accuracy of affect detection models, as well as for gaining a more comprehensive understanding of the full range of students' emotional experiences in learning environments.
    - **Explore additional emotions**: The current literature has predominantly focused on four affective states—boredom, confusion, engagement, and frustration. This emphasis is largely due to these emotions being the most frequently detected in previous research on learning environments, as highlighted by S. D'Mello and Calvo (2013). In their analysis, they found that these four emotions occurred at five times the rate of basic emotions like anger, sadness, or disgust across various tasks and methodologies. Future research should investigate whether other emotions, such as delight, anxiety, or interest, might also be relevant in different types of CBLEs. Additionally, it would be valuable to explore whether the occurrence and frequency of these emotions change depending on the learning environment or content. By expanding the emotional spectrum beyond these frequently detected states, researchers can gain a more comprehensive understanding of students' emotional experiences and responses in diverse educational contexts.

2. **Enhancing Model Development Practices and Methods**: The development of sensor-free affect detection models has largely been driven by a small number of research groups, often using data collection methods based on online classroom observation. However, there is a need to diversify and refine these methods to improve model accuracy and applicability. Future research could focus on:

    - **Compare the Accuracy of Various Data Collection Techniques**: Current research heavily relies on online classroom observation for data collection. Future studies should explore and compare the accuracy of different techniques to determine the most effective approaches across diverse learning environments and contexts, potentially leading to more robust and generalizable models.
    - **Determine the Ideal Granularity of Duration**: Determining the ideal granularity of duration and assessing whether this parameter depends on the type of CBLE is crucial. Future research should explore if a better grain level exists and whether it depends on the learners or CBLEs.
    - **Shared Database of Action Logs and Emotion Labels**: The field of sensor-free affect detection could significantly benefit from a shared database of action logs and emotion labels. Such a resource would enable researchers to develop and compare their models using the same dataset.
    - **Open Source Code**: In addition to a shared database, making the source code of these models publicly available in libraries would allow for greater transparency, reproducibility, and collaboration in the field. Future research should consider ways to facilitate this.

3. **Integrating Models into CBLEs**: The practical application of these models in real learning environments is a crucial next step. This would allow for immediate intervention and adaptation based on the detected emotions, potentially optimizing learning outcomes. Future research directions in this area could include:

    - **Integrate Models into CBLEs for Real-Time Detection**: Future research could focus on integrating these models into CBLEs and performing real-time detection of students' emotions. This would allow for immediate intervention and adaptation based on the detected emotions, potentially optimizing learning outcomes. In our SLR, we were able to find a unique work that integrated the model into a CBLE.
    - **Provide Meaningful Interventions Based on Detected Emotions**: Once emotions are detected in real-time, the next challenge is to provide meaningful interventions. Future research could explore how to best respond to detected emotions to enhance the learning experience.
    - **Understand the Impact of Emotions on Learning**: While it is well-documented that emotions influence learning, there is still much to understand about this relationship. Future research could delve deeper into how specific emotions impact different aspects of learning, such as engagement, motivation, and learning outcomes.

In conclusion, while substantial progress has been made in the field of sensor-free affect detection, there are numerous avenues for future research. By addressing these areas, we can move closer to the goal of creating CBLEs that are truly responsive to students' emotional states, thereby optimizing the learning process.

8 Limitations and Threats to validity

This systematic review, while comprehensive, has several limitations that should be acknowledged. First, the scope of the review was limited to studies published in English. This language restriction may have led to the omission of relevant studies published in other languages. Second, the review relied on specific databases for the literature search. While these databases are widely used and respected, there may be relevant studies in other databases or grey literature that were not included in this review. Third, the review process was subject to potential reviewer bias. Despite efforts to minimize this through independent reviews and consensus discussions, the subjective nature of study selection and data extraction processes may have influenced the results of the review. Fourth, the review did not conduct a formal assessment of the quality or risk of bias in the individual studies included. It does limit the ability to assess the validity and reliability of the findings of the individual studies. Fifth, the review did not include a meta-analysis due to the heterogeneity of the studies in terms of methodologies and measures used. This limits the ability to quantitatively synthesize the results and draw definitive conclusions.

Future systematic reviews could address these limitations by including studies in other languages, searching additional databases, conducting a formal quality assessment, and regularly updating the review to include new research.

Declaration of Generative AI Software tools in the writing process

In the development of this manuscript, the authors utilized the ChatGPT service, which is based on the GPT-4o language model, specifically for proofreading purposes. Following the application of this tool, the authors meticulously revised and refined the content as necessary. It is imperative to highlight that the authors assume full responsibility for the final content of the publication.

Acknowledgements

This work was supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001, by FAPERGS (Process 17/2551-0001203-8), and by CNPq (Process 306005/2020-4).

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. *International Conference on Intelligent Tutoring Systems (ITS)*, 227–239. https://doi.org/10.1007/978-3-540-30139-4_22 [GS Search].

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, *5*(3), 176–204. [GS Search].

Andres, J. M. A. L., Ocumpaugh, J., Baker, R., Slater, S., Paquette, L., Jiang, Y., Karumbaiah, S., Bosch, N., Munshi, A., Moore, A., & Biswas, G. (2019). Affect sequences and learning in betty's brain. *International Conference on Learning Analytics and Knowledge (LAK)*, 383–390. https://doi.org/10.1145/3303772.3303807 [GS Search].

Arroyo, I., & et al. (2009). Emotion sensors go to school. *International Conference on Artificial Intelligence in Education (AIED)*, *200*, 17–24. https://doi.org/10.3233/978-1-60750-028-5-17 [GS Search].

Arroyo, I., Muldner, K., Schultz, S., Burleson, W., Wixon, N., & Woolf, B. P. (2016). Addressing affective states with empathy and growth mindset. *24th ACM Conference on User Modeling, Adaptation and Personalisation (UMAP)*, 314–319. https://doi.org/10.1007/978-3-319-39583-8_35 [GS Search].

Azevedo, R., & Aleven, V. (2013). *International handbook of metacognition and learning technologies* (Vol. 26). Springer. [GS Search].

Baker, R., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-task behavior in the cognitive tutor classroom: When students game the system. *ACM CHI 2004: Computer-Human Interaction*, 383–390. https://doi.org/10.1145/985692.985741 [GS Search].

Baker, R. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of the SIGCHI conference on Human factors in computing systems (ACM CHI)*, 1059–1068. https://doi.org/10.1145/1240624.1240785 [GS Search].

Baker, R., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. (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 [GS Search].

Baker, R., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G. W., Ocumpaugh, J., & Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra: Population validity for educational data mining. *British Journal of Educational Technology*, *45*(3), 487–501. https://doi.org/10.1111/bjet.12156 [GS Search].

Baker, R., & Ocumpaugh, J. (2014). Interaction-based affect detection in educational software. In R. A. Calvo, S. K. D'Mello, J. Gratch, & A. Kappas (Eds.), *The oxford handbook of affective computing*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.009 [GS Search].

Baker, R., Ocumpaugh, J., Gowda, S. M., Kamarainen, A. M., & Metcalf, S. J. (2014). Extending log-based affect detection to a multi-user virtual environment for science. *International Conference on User Modeling, Adaptation, and Personalization*, 290–300. https://doi.org/10.1007/978-3-319-08786-3_25 [GS Search].

Bloom, B. S. (1977). *Human characteristics and school learning.* (Vol. 10). McGraw-Hill. https://doi.org/10.2307/1478496 [GS Search].

Bosch, N., Chen, H., D'Mello, S., Baker, R., & Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 267–274. https://doi.org/10.1145/2818346.2820739 [GS Search].

Bosch, N., & D'Mello, S. (2017). The affective experience of novice computer programmers. *IJAIED*, *27*(1), 181–206. https://doi.org/10.1007/978-3-030-23204-7_25 [GS Search].

Bosch, N., D'Mello, S., & Mills, C. (2013). What emotions do novices experience during their first computer programming learning session? *International Conference on Artificial Intelligence in Education (AIED)*, 11–20. https://doi.org/10.1007/978-3-642-39112-5_2 [GS Search].

Botelho, A. F., Baker, R., & Heffernan, N. T. (2017). Improving sensor-free affect detection using deep learning. *International Conference on Artificial Intelligence in Education (AIED)*, 40–51. https://doi.org/10.1007/978-3-319-61425-0_4 [GS Search].

Botelho, A. F., Baker, R., & Heffernan, N. T. (2019). Machine-learned or expert-engineered features? exploring feature engineering methods in detectors of student behavior and affect. *International Conference on Educational Data Mining (ICEDM)*. [GS Search].

Calvo, R. A., & D'Mello, S. K. (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 [GS Search].

Carroll, J. B. (1963). A model of school learning. *Teachers college record*, *3*(8), 155–167. https://doi.org/10.1177/016146816306400801 [GS Search].

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357. https://doi.org/10.1613/jair.953 [GS Search].

Cocea, M., Hershkovitz, A., & Baker, R. (2009). The impact of off-task and gaming behaviors on learning: Immediate or aggregate? *International Conference on Artificial Intelligence in Education (AIED)*, 507–514. https://doi.org/10.3233/978-1-60750-028-5-507 [GS Search].

Cocea, M., & Weibelzahl, S. (2010). Disengagement detection in online learning: Validation studies and perspectives. *IEEE transactions on learning technologies*, *4*(2), 114–124. https://doi.org/10.1109/tlt.2010.14 [GS Search].

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of Eeducational Media*, *29*(3), 241–250. https://doi.org/10.1080/1358165042000283101 [GS Search].

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal performance*. Harper; Row. [GS Search].

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

de Morais, F., & Jaques, P. A. (2023). The dynamics of brazilian students' emotions in digital learning systems: Investigating the interaction between gender and emotions duration. *International Journal of Artificial Intelligence in Education (IJAIED)*, ((in press)). https://doi.org/10.1007/s40593-023-00339-0 [GS Search].

de Morais, F., Kautzmann, T. R., Bittencourt, I. I., & Jaques, P. A. (2019). EmAP-ML: A protocol of emotions and behaviors annotation for machine learning labels. *European Conference for Technology-Enhanced Learning (EC-TEL)*. https://doi.org/10.1007/978-3-030-29736-7_37 [GS Search].

D'Mello, S. (2011). Dynamical emotions: Bodily dynamics of affect during problem solving. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33). [GS Search].

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. https://doi.org/10.1037/a0032674 [GS Search].

D'Mello, S., & Calvo, R. A. (2013). Beyond the basic emotions: What should affective computing compute? *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 2287–2294. https://doi.org/10.1145/2468356.2468751 [GS Search].

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001 [GS Search].

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. https://doi.org/https://doi.org/10.1016/j.learninstruc.2012.05.003 [GS Search].

D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., Perkins, L., & Graesser, A. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (pp. 245–254, Vol. 6094 LNCS). https://doi.org/10.1007/978-3-642-13388-6_29 [GS Search].

D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. (2006). Predicting affective states expressed through an emote-aloud procedure from autotutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education (IJAIED)*, 3–28. [GS Search].

Finn, J. D. (1989). Withdrawing from school. *Review of educational research*, *59*(2), 117–142. https://doi.org/10.3102/00346543059002117 [GS Search].

Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology*, *2*(3), 300–319. https://doi.org/10.1037/1089-2680.2.3.300 [GS Search].

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics—a "hopeless" issue? a control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*(4), 497–514. https://doi.org/10.1007/BF03173468 [GS Search].

Graesser, A., & D'Mello, S. (2012). Emotions during the learning of difficult material. In *Psychology of learning and motivation* (pp. 183–225, Vol. 57). Elsevier. https://doi.org/10.1016/b978-0-12-394293-7.00005-4 [GS Search].

Graesser, A., D'Mello, S., & Strain, A. C. (2014). Emotions in advanced learning technologies. In *International handbook of emotions in education* (pp. 483–503). Routledge. [GS Search].

Graesser, A., & D'Mello, S. K. (2011). Theoretical perspectives on affect and deep learning. In *New perspectives on affect and learning technologies* (pp. 11–21). Springer. https://doi.org/10.1007/978-1-4419-9625-1_2 [GS Search].

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77. https://doi.org/10.3102/00346543058001047 [GS Search].

Henderson, N., Rowe, J., Paquette, L., Baker, R., & Lester, J. (2020). Improving affect detection in game-based learning with multimodal data fusion. *International Conference on Artificial*

Intelligence in Education (AIED), 228–239. https://doi.org/10.1007/978-3-030-52237-7_19 [GS Search].

Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review*, *1*(2), 69–82. https://doi.org/10.1016/j.edurev.2006.09.001 [GS Search].

Hutt, S., Grafsgaard, J. F., & D'Mello, S. (2019). Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300726 [GS Search].

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly*, *14*(3), 299–324. https://doi.org/10.1111/j.1471-6402.1990.tb00022.x [GS Search].

Jensen, E., Hutt, S., & D'Mello, S. K. (2019). Generalizability of sensor-free affect detection models in a longitudinal dataset of tens of thousands of students. *International Educational Data Mining Society*. [GS Search].

Jiang, Y., Bosch, N., Baker, R., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., Moore, A. L., & Biswas, G. (2018). Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? *AIED*, 198–211. https://doi.org/10.1007/978-3-319-93843-1_15 [GS Search].

Kai, S., Paquette, L., Baker, R., 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. *International Conference on Educational Data Mining (ICEDM)*, 44–53. https://doi.org/10.1007/978-3-319-19773-9_5 [GS Search].

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 [GS Search].

Lee, D. M. C., Rodrigo, M. M. T., d Baker, R., Sugay, J. O., & Coronel, A. (2011). Exploring the relationship between novice programmer confusion and achievement. *International Conference on Affective Computing and Intelligent Interaction*, 175–184. https://doi.org/10.1007/978-3-642-24600-5_21 [GS Search].

Lee, S. W., Kelly, K. E., & Nyre, J. E. (1999). Preliminary report on the relation of students' on-task behavior with completion of school work. *Psychological Reports*, *84*(1), 267–272. https://doi.org/10.2466/pr0.1999.84.1.267 [GS Search].

Lehman, B., D'Mello, S., & Graesser, A. (2012). Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, *15*(3), 184–194. https://doi.org/10.1016/j.iheduc.2012.01.002 [GS Search].

Lehman, B., D'Mello, S., & Graesser, A. (2012). Interventions to regulate confusion during learning. *International Conference on Intelligent Tutoring Systems (ITS)*, 576–578. https://doi.org/10.1007/978-3-642-30950-2_75 [GS Search].

Lehman, B., & et al. (2008). What are you feeling? investigating student affective states during expert human tutoring sessions. *ITS*, 50–59. https://doi.org/10.1007/978-3-540-69132-7_10 [GS Search].

Litman, D., & Forbes-Riley, K. (2014). Evaluating a spoken dialogue system that detects and adapts to user affective states. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 181–185. [GS Search].

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of educational psychology*, *88*(2), 203. https://doi.org/10.1037/0022-0663.88.2.203 [GS Search].

Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill. [GS Search].

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487–501. https://doi.org/10.1111/bjet.12156 [GS Search].

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2015). Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*. [GS Search].

Paquette, L., Baker, R., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., & Kauffman-Rogoff, Z. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. *ITS*, 1–10. https://doi.org/10.1007/978-3-319-07221-0_1 [GS Search].

Paquette, L., Rowe, J., Baker, R., Mott, B., Lester, J., DeFalco, J., Brawner, K., Sottilare, R., & Georgoulas, V. (2016). Sensor-free or sensor-full: A comparison of data modalities in multi-channel affect detection. *International Conference on Educational Data Mining (EDM)*. [GS Search].

Pardos, Z. A., Baker, R., San Pedro, M. O., 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. *JLA*, *1*(1), 107–128. https://doi.org/10.18608/jla.2014.11.6 [GS Search].

Patrick, B. C., Skinner, E. A., & Connell, J. P. (1993). What motivates children's behavior and emotion? joint effects of perceived control and autonomy in the academic domain. *Journal of Personality and social Psychology*, *65*(4), 781. https://doi.org/10.1037/0022-3514.65.4.781 [GS Search].

Pedro, M. O., Baker, R., Bowers, A., & Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Educational Data Mining 2013*. [GS Search].

Pekrun, R. (2007). Emotions in students' scholastic development. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 553–610). Springer. https://doi.org/10.1007/1-4020-5742-3_13 [GS Search].

Pekrun, R., et al. (2014). *Emotions and learning* (Vol. 24). International Academy of Education (IAE) Geneva, Switzerland. [GS Search].

Pekrun, R. (2016). Academic emotions. *Handbook of motivation at school*, *2*, 120–144. https://doi.org/10.4324/9781138609877-ree210-1 [GS Search].

Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of educational Psychology*, *98*(3), 583. https://doi.org/10.1037/0022-0663.98.3.583 [GS Search].

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002a). Positive emotions in education. In *Beyond coping: Meeting goals, visions, and challenges* (pp. 149–173). Oxford University Press. https://doi.org/10.1093/med:psych/9780198508144.003.0008 [GS Search].

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002b). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research.

Educational psychologist, *37*(2), 91–105. https://doi.org/10.4324/9781410608628-4 [GS Search].

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement* (pp. 259–282). Springer. https://doi.org/10.1007/978-3-031-07853-8_6 [GS Search].

Perkins, H. V. (1965). Classroom behavior and underachievement. *American Educational Research Journal*, *2*(1), 1–12. https://doi.org/10.3102/00028312002001001 [GS Search].

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic mapping studies in software engineering. *EASE*, *8*, 68–77. https://doi.org/10.14236/ewic/ease2008.8 [GS Search].

Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., & et al. (2013). Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education (IJAIED)*, *22*(3), 107–140. https://doi.org/10.1007/s40593-023-00346-1 [GS Search].

Reis, H., Alvares, D., Jaques, P., & Isotani, S. (2018). Analysis of permanence time in emotional states: A case study using educational software. *Intelligent Tutoring Systems*, 180–190. https://doi.org/10.1007/978-3-319-91464-0_18 [GS Search].

Rodrigo, M. M. T., Baker, R., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., Lim, S. A. L., Pascua, S. A., Sugay, J. O., & Tabanao, E. S. (2009). Affective and behavioral predictors of novice programmer achievement. *Proceedings of the 14th annual ACM SIGCSE conference on Innovation and technology in computer science education*, 156–160. https://doi.org/10.1145/1595496.1562929 [GS Search].

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295x.110.1.145 [GS Search].

Sabourin, J., Mott, B., & Lester, J. C. (2011). Modeling learner affect with theoretically grounded dynamic bayesian networks. *International Conference on Affective Computing and Intelligent Interaction*, 286–295. https://doi.org/10.1007/978-3-642-24600-5_32 [GS Search].

Salmeron-Majadas, S., Santos, O. C., & Boticario, J. G. (2014). An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context. *Procedia Computer Science*. [GS Search].

Scherer, K. R. (2000). Psychological models of emotion. *The neuropsychology of emotion*, *137*(3), 137–162. [GS Search].

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216 [GS Search].

Smeets, R., Broaekman, F., & Bouwers, E. (2019). Affect detection in home-based educational software for young children. *International Journal of Educational Research*, *107*, 101746. https://doi.org/10.1016/j.ijer.2021.101746 [GS Search].

Tiam-Lee, T. J., & Sumi, K. (2019). Analysis and prediction of student emotions while doing programming exercises. In *International conference on intelligent tutoring systems (its)* (pp. 24–33). Springer. https://doi.org/10.1007/978-3-030-22244-4_4 [GS Search].

Vea, L., Rodrigo, M., et al. (2016). Modeling negative affect detector of novice programming students using keyboard dynamics and mouse behavior. *Pacific Rim International Conference on Artificial Intelligence*, 127–138. https://doi.org/10.1007/978-3-319-60675-0_11 [GS Search].

Verduyn, P., Delaveau, P., Rotgé, J.-Y., Fossati, P., & Mechelen, I. V. (2015). Determinants of emotion duration and underlying psychological and neural mechanisms. *Emotion Review*, *7*(4), 330–335. https://doi.org/10.1177/1754073915590618 [GS Search].

Verduyn, P., & Lavrijsen, S. (2015). Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion*, *39*, 119–127. https://doi.org/10.1007/s11031-014-9445-y [GS Search].

Verduyn, P., Van Mechelen, I., & Tuerlinckx, F. (2011). The relation between event processing and the duration of emotional experience. *Emotion*, *11*(1), 20–28. https://doi.org/10.1037/a0021239 [GS Search].

Wang, Y., Heffernan, N. T., & Heffernan, C. (2015). Towards better affect detectors: Effect of missing skills, class features and common wrong answers. *International Conference on Learning Analytics and Knowledge (LAK)*, 31–35. https://doi.org/10.1145/2723576.2723618 [GS Search].

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). *Data mining: Practical machine learning tools and techniques* (Vol. 2). Elsevier. https://doi.org/10.1016/b978-0-12-374856-0.00015-8 [GS Search].

Wixon, M., & Arroyo, I. (2014). When the question is part of the answer: Examining the impact of emotion self-reports on student emotion. *International Conference on User Modeling, Adaptation, and Personalization*, 471–477. https://doi.org/10.1007/978-3-319-08786-3_42 [GS Search].

Yang, D., Kraut, R. E., & Rose, C. P. (2016). Exploring the effect of student confusion in massive open online courses. *Journal of Educational Data Mining*, *8*(1), 52–83. [GS Search].