# Scalable Facial Behaviors Sensing to Detect States of Flow in Online Learning Through A Privacy-Preserving Framework

While active engagement is critical to effective learning, the immense amount of potential distractions, either online or in-person, in home environments prevent students from reaching the pinnacle of engagement. The ability to recognize whether or not an individual is in flow provides further insight into their learning experiences, which can serve as a useful tool in online learning. In this proof-of-concept study, we developed a privacy-preserving framework and collected multimodal facial behavioral features (action units, head pose, and eye aspect ratio) from college students, who self-reported their level of flow along with other affective states while participating in an online coding activity. Using this data, we developed a machine learning model that detects states of flow with an accuracy of 92.65% (F1 = 86.95) and highlight the most important facial behavior markers contributing to model performance. Not only does this study demonstrate the potential of facial behavior markers using passive sensing and machine learning for identifying a student's flow level in a naturalistic online learning setting, but it opens the door towards just in time interventions depending on that level.

CCS Concepts: • **Human-centered computing**; • **Human Computer Interaction (HCI)**; • **Empirical studies in HCI**;

Additional Key Words and Phrases: Flow, Facial Behavior Markers, Passive Sensing, Machine Learning, Light Gradient Boosting Machine (LGBM)

**ACM Reference Format:**
. 2021. Scalable Facial Behaviors Sensing to Detect States of Flow in Online Learning Through A Privacy-Preserving Framework. 1, 1 (October 2021), 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Flow is described as a state of complete immersion, in which one's focus is so concentrated on the task at hand, time stops and the rest of the world falls away [9]. A person is not simply in the process of completing a task, they know exactly how to proceed and will not (or cannot) stop until they achieve their goal. As it transcends many standard mental states, flow is understandably difficult to achieve. Without the necessary characteristics, such as an interesting, complete-able task with clear goals, one can easily fall out of flow, into a more negative mental state. It is widely accepted that, in order to achieve a flow state, the perceived challenge from an activity must equal an individual's perceived skills. Without this equilibrium, an individual falls into less ideal affective states, such as anxiety or boredom [8, 35]. Nevertheless, flow is considered the optimal mental state, and those who experience flow will most likely undergo a more satisfying, engaging, and, overall, better experience.

One of the most notable applications of flow theory has been in the video game context. User engagement and state of immersion can be used to evaluate successful game design [22, 34], so researchers are interested in detecting and modeling the flow state of video game users to optimize the success and popularity of a game. Successful studies have utilized physiological signals collected with extraneous devices, such as heart rate variability or electrodermal activity, to successfully detect a participant's flow while they are playing video games

Author's address:


XXXX-XXXX/2021/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[30, 34]. Similar studies in game-based learning have been successful in detecting flow, in this case by using data logs [38] or psychometric surveys [21].

However, few studies have attempted to identify and/or evaluate flow in the purely educational context, and, most of the studies only attempt to detect flow after students have completed an entire course. Both Semerci et al., and Lynch et al.'s studies attempt to detect a student's flow state after completing a course on an e-learning platform or system [29, 41]. Furthermore, each study uses the characteristics of a student's interaction with a system to detect flow, as compared to studies in the video game context [22, 34], which use characteristics and features of the students themselves. In fact, as far as we can tell, there is a lack of studies attempting to detect states of flow in the online learning context by incorporating the physical characteristics of students.

The only comparable studies are those who have utilized the physical characteristics of students to detect other affective states in online learning. Many existing affective state detection studies incorporate a participant's facial muscle action units (AUs), facial expression (FE), head posture (HP), and/or eye aspect ratio (EAR), gaze and other eye features in their detection process [2, 4, 5, 22, 25, 36, 45]. However, either while extracting of features or determining ground truths, almost all affective state (and even flow) studies rely on either participant recordings, extraneous devices, or both. Although processes using participant recordings have been successful, the existence and continued use of participant recordings increases participant privacy concerns.

Affective state detection studies can differ based on the methods used to report ground truth values. Studies can either use self-reports from those being observed or observational evaluations from an external observer. While the previously mentioned flow detection studies all incorporate subjective self-reports, many other affective state detection studies incorporate observer evaluations [4, 6, 25, 45]. Although some studies that incorporate observer evaluations have been successful, we believe that there are gaps in these methods. Observer evaluations rely on an outside human perspective, which may be biased because of the inherent bias in human views. Bias in evaluations leads to bias in data, which leads to bias in any machine learning models trained with that data. Subjective student reports provide a more accurate representation of a student's actual affective state.

To fill in the gaps in existing research, we propose a proof-of-concept study to detect a student's state of flow in the real-world context, while they are participating in a programming activity. We developed a novel facial behavior sensing framework (the FacePsy framework), which unobtrusively captures facial behaviors while students work. The FacePsy framework does not require extra devices (e.g. eye-trackers) or privacy-invasive video recordings to monitor students' behaviors, and instructors are not required to observe and annotate student behavior. Therefore, the goal of our study is to answer the following questions:

- RQ1: Can the facial behavior features collected in real-world contexts through the FacePsy framework be used to observe relationships between the states of flow, and can they observe the differences between these states?
- RQ2: Can these key features be used to build a machine learning model for identifying students' subjective reported states of flow (high vs. low)?

Our study models students' perceived flow state, leveraging unobtrusively collected facial behavior markers (AUs, HP, FE, and EAR). This proof-of-concept study develops novel approaches in identifying a self-reported flow state from college students by developing a facial behavior sensing framework that extracts features directly while running instead of extracting from recorded facial images, the process long used by previous studies [4, 25, 33, 45].

Building upon the limitations of existing studies, our contributions are as follows: (1) Our study first proposes a state of flow detection model trained, tested, and validated using students' subjective self-reports. (2) We advance objective behavioral markers that identify states of flow by integrating a prominent set of facial movements, poses, and facial expressions.

In the following sections, we review previous studies on flow, flow detection models, and different types of objective markers measuring flow states. Next, we introduce our framework, describe our method for data collection, data pre-processing, and modeling states of flow: low- and high-flow state. We conclude with a discussion of the implications and contributions of our model development.

## 2 BACKGROUND AND RELATED WORK

### 2.1 The Relationship between Flow, Anxiety, and Boredom

Csikszentmihalyi's flow theory refers to a state of mind where one is completely immersed in an activity [9]. As seen in Figure 1, a state of flow is most likely achieved when a participant believes the challenge of an activity matches their skill levels [35]. When the perceived level of challenge rises above a participant's skill level, they experience anxiety or frustration, and, when an activity does not challenge a participant to the height of their skill levels, they experience boredom. While detecting and examining the state of flow, simultaneously detecting and examining the states of boredom and anxiety or frustration may provide further insight into a participant's learning experience.

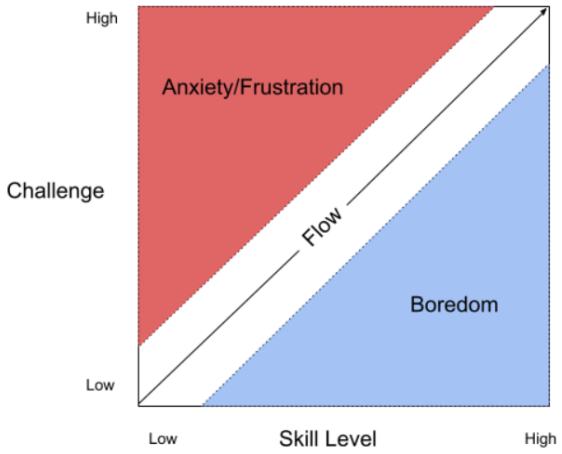


Fig. 1. Flow State Model (Adapted from [8])

Studies have successfully examined the relationship between and the characteristics of a participant's measured affective states, even if the purpose of their research centered around a specific state. For example, Lee et al. incorporated flow theory in their research into an automatic flow detector. They created models that successfully distinguish whether a student is in a state of flow, boredom, or frustration, based on certain features of an online intelligent tutoring system for a linear algebra course. The researcher's conclusions follow the affective state relationships recognized in flow theory and support the use of affective state detectors in detecting the conditions of flow [27]. Comparatively, Baker et al.'s research sought to examine the incidence, persistence, and impact of a student's affective state in three unique online learning environments. Confusion and engaged concentration (a component of flow) were the most prominent affective states within the environments, and boredom was the most persistent state. Boredom was also the only state associated with 'gaming the system,' and the researchers noted the difficulty of transitioning from boredom to another state [3]. It is important to note that both of these studies, while related to online learning, both involve intelligent tutoring systems. It is unclear if the relationship

between the states, as described in flow theory, have been explored while detecting affective states in other educational contexts.

## 2.2 Flow in the Educational Context

The application of flow detection in education, either in-person or online learning, appears to be relatively new. To the best of our knowledge, the few studies found related to flow detection in the learning context were all conducted relatively recently. Although these studies have reported successful results in detecting a student's flow, research in this specific field lacks the variety of methodology and contexts seen in similar fields. Most of the studies found attempt to measure a student's flow after completing an entire online course through a specific system. For example, Semerci and Goularas used interaction data from an e-learning platform, quiz results, and surveys to calculate activity and performance, and plotted these calculations to evaluate the flow state of students taking an online course [41]. Similarly, Lynch and Ghergulescu's determined whether students experienced flow, anxiety, and boredom while using the Adaptemy e-learning system, by finding the relationship between each student's self-estimated ability, reported after using Adaptemy for a math course, and the difficulty level generated by the system itself [29]. Both studies collected features from students that related to a their interaction with some kind of online learning system or platform. It seems that, in this context, little to no studies attempt to detect flow while a student participates in an individual activity. We were also unable to find any studies that implemented the physical features of particiapants (either physiological or facial) into detecting flow in a purely educational setting.

## 2.3 Flow Detection in Video Games and Game-based Learning

Gaming is a prevalent context for flow or engagement detection. Game developers want their games to bring about a flow state in users in order to increase a game's popularity and success [22, 34]. There have been successful studies that use a participant's physiological signals [30, 34], physical characteristics, or a combination of the two [22, 46] to recognize a state of flow or engagement. When incorporating physiological signals into emotion recognition, extra devices are most always necessary. Each of the previously mentioned studies incorporate some kind of extra device. Maier et al. equipped participants with an Empatica E4 wrist device that captures heart rate, heart rate variability, blood volume pulse (BVP), electrodermal activity (EDA) and skin temperature, and used the raw BVP and/or EDA data in their DeepFlow detection model [30]. Huynh et al. used a photoplethysmography sensor to measure heart rate variability, an electrodermal activity sensor to measure skin conductance response and phasic component series, a touchscreen sensor to measure touch events, and a depth camera to measure upper-body movement in their EngageMon model, which successfully classified the level of participant engagement, but was unsuccessful in measuring engagement in real-time [22]. Michailidis et al. used electrocardiography recordings, combined with participant reports to successfully classify data as either high or low flow [34]. And, Yang et al. used a combination of peripheral physiological signals (electrocardiogram, electrodermal, electromyography, respiration, and body movement), facial recordings, game screen recordings, and other meta information (e.g. player skill level, game difficulty, and game resulting score) to create a successful emotion recognition model [46].

Because flow in video games is a widely recognized subset of flow research, there has been a surge in studies related to game-based learning, many of which attempt to apply game-based flow theory to learning. Game-based learning refers to the use of game elements and characteristics, like points, badges, progress bars, and profile development in the educational context [38]. Pastushenko et al. proposed to analyze the characteristics of data logs from participants who completed a game-based assignment (time in the system, time to finish an activity, proportion of correct activities, etc.), combined with calculated flow experience to evaluate the influence of the detected flow state on learning outcomes [38]. Similarly, Hamari et al. investigated the impact of flow, engagement,

and immersion in game-based learning and found that, based on a psychometric survey, educational games increase student engagement in an activity, which in turn positively effects learning [21].

Despite several successful studies in the video game context, flow detection and, to a further extent, all emotion recognition still has room for improvement. Many studies that aim to detect and/or model flow or engagement in the video game context require extra devices, which may interfere with a participant's ability to play. For example, while Maier et al.'s DeepFlow Framework is comparable to the FacePsy framework, it relies on wearable devices to extract physiological signals. The notion of extraneous devices also raises questions about the feasibility of applying these studies into the real world, especially to online learning. It would be extremely difficult to ensure that participants are correctly wearing and using extraneous devices, which puts the quality of collected data into question.

## 2.4 Existing Methods in Affective State Detection

While there does not appear to be many studies interested in detecting a student's flow state in online learning, there is an abundance of studies that attempt to detect engagement in online learning. Flow and engagement are generally similar, but they can not be considered as the same states. Engagement or engaged concentration can be considered one of the components necessary to achieve a flow state [9]. We belive the methods used in the detection of other affective states, most notably engagement, can provide further insight into and even potentially be applied to flow state detection.

Existing methods to detect student engagement in online learning can be classified according to the extent of the student's participation in said method. The three possible categories of detection are Manual engagement detection, Semi-automatic engagement detection, and Automatic engagement detection. Automatic engagement detection, which automatically extracts specific features from either data collected by either image sensors, physiological or neurological sensors, or monitored learner activity [10], is one of the most popular categories for detection and is the most relevant to this study. Automatic detection itself can be further divided into three sub-categories: log-file analysis, sensor data analysis, and computer vision based methods. In log-file analysis, researchers analyze participant's actions, which are stored in log files, similarly to Pastushenko et al., Lee et al., and Semerci and Goularas flow detection studies [26, 38, 41]. In sensor data analysis, researchers implement physiological and/or neurological sensors into their studies and use the readings to measure engagement [10]. Similar methods have also been applied to flow detection [30, 34, 39]. To the best of our knowledge, only computer vision based methods, which measure participant engagement by analyzing unobtrusively detected facial and positional cues, have yet to been applied to flow detection.

The most commonly used modalities in computer vision based engagement detection are gestures and postures, eye movement, and facial expressions [10]. Facial detection is easily the most prominent modality of computer vision based engagement detection, and its methods can categorized as either part-based or appearance-based. Part-based methods are techniques that analyze different parts of a participant's face, such as the eyes, mouth, and chin, as compared to appearance-based methods, where features from whole-face regions are analyzed [10]. Part-based methods usually analyze specific or combinations of specific facial muscle movements, known as action units (AUs), as derived from the Facial Action Coding System [13, 16].

Existing research has looked extensively into mapping AUs to specific facial expressions, such as anger, sadness, and happiness [31] and many have used a combination of AUs with other modalities, such as head pose (HP) or eye aspect ratio (EAR) to detect engagement. The specific AUs chosen for observation depend on the study, but some of the most prominent AUs seen in affective state detection include include AU01 [4, 19], AU02 [19, 20], AU04 [4, 12, 19, 20, 28], and AU14 [4, 19]. Grafsgaard et al.'s [19] research into automatically recognizing facial indicators of frustration was one of the first studies that applied AU recognition (specifically AUs 1, 2, 4, 7, and 14) to the detection of affective states, including engagement, although the only significant

correlations involved participant-reported frustration. Whitehill et al. [45] and Bosch et al. [6] successfully created an engagement detector with comparable accuracy to a human observer and used a variety of machine learning classifiers to successfully detect six affective states (Delighted, Off Task, Engaged, Frustrated, Confused, and Bored), respectively, both incorporating facial features and head positions extracted from videos of students. Notably, Aslan et al. successfully detected engagement during 1:1 learning using a combination of student body postures, facial expressions, and eye gaze, using observer evaluations as ground truths. The researchers used the top 25 features correlated with engagement to train a Decision Tree classifier and achieved an accuracy of about 85% [2].

While these studies have been successful in utilizing facial features to detect online student engagement in real-life settings, they are not perfect. The process of facial feature extraction is over-complicated and most always occurs after a participant has finished the experiment. Extracting features from videos of participants after the experimental portion of a study has concluded not only adds a privacy risk to participants, but the entire act of recording, storing, and analyzing videos of participants increases the potential for a privacy violation. Again, the use of extra devices in online learning provides a risk to the quality of data collected.

### 2.5 The Use of Observer Evaluations vs. Self-reports in Recording Ground Truths

An important distinction in any affective state detection is the use of observer evaluations compared to the use of self-reports to record the ground truth of an affective state. In self-reporting, a participant responds to either an open or closed prompt with their personal belief, behavior, or answer. In observational checklists, an individual fills out a questionnaire asking for their qualitative or quantitative opinions on another individual. Although each of the previously mentioned flow detection studys use a participant's self-reported flow as a ground truth, other affective state detection studies are generally evenly split between self-reported ground truths and observer evaluated ground truths.

Interestingly, in the majority of computer vision-based affective state detection studies found, observer evaluated ground truths were used [2, 4, 5, 18, 25, 45], with only three using self-reports [6, 19, 36]. Almost every study that incorporated log-file analysis or sensor data analysis used a participant's self-report as a ground truth, with only one [3] using observer evaluations.

The specific methods in either ground truth reporting method vary depending on the study. When it comes to self-reporting, studies may use a previously validated questionnaire [30, 34, 38, 39] or ask participants about the state directly [11, 36, 46] or indirectly [17, 19, 21, 22, 26], usually using a scale. For example, participants in the Yang et al. study ranked their happiness, frustration, proudness, curiosity, anger, fear, boredom, and sadness on a scale from -3 to 3 [46], while participants in the Hamari et. al study answered questions like "How interesting was the game? Do you with you were doing something else?" [21]. In most observer evaluations, observers view a video of a participant and either choose which mental state is the most prominent [3, 18, 33] or rank the intensity of a mental state on a given scale [2, 4, 45].

## 3 THE FACEPSY FRAMEWORK

The system architecture and implementation of our facial behavior sensing framework is illustrated in Figure 2. The system has two main interfaces: (a) *Student interface*, and (b) *Researcher/Teacher interface*. The *Student interface* implements study protocol and data collection setups. The *Student interface* adopts AI based models to replace the need for human annotators within a study protocol. The collected data from the *Student interface* is transferred to an intermediate storage server. We provided various APIs to consume the stored data from different applications and use cases. In the following sections, we review these key interfaces in greater detail, focusing on the adoption of AI in our sensing framework.

### 3.1 Our Facial Behavior Sensing Framework

The FacePsy framework (Figure 2) implements state-of-the-art facial behavior detection modules, such as Action Unit (AU) detection, Facial expression recognition, Facial Landmark detection, and Head pose estimation, which are all essential for modeling complex mental states [14]. To circumvent traditional study requirements, our application is designed to run on commodity hardware with limited computation and memory allocation. It leverages state-of-the-art technologies, including the dlib face and shape detector for face detection and 68 landmark detection tasks [24]. For AU detection we use [15], and, for facial expression recognition, we leverage existing model architecture and weights from [42]. The combined pipeline of all these tasks builds to an in-device feature extraction module, which runs at 7 FPS (Frames Per Second). For collecting participant surveys, we use Google forms. Our application generates a unique survey link for each session, with pre-filled participation identifiers and session identifiers. The application implements three controls (Fig. 2) in the GUI: (a) the Start button: to start data collection, (b) the Stop button: to stop data collection, and (c) the Survey button: to open the survey link in a browser.

Upon installation, the FacePsy framework generates a unique participation identifier for each individual. The unique identifier, with a length of 8, is generated using shortuuid [43]. Upon clicking the 'Start button' on the FacePsy framework GUI, the application loads feature extraction modules and initiates a connection to the computer's webcam. Each camera frame from the webcam is then processed through the pipeline to extract low level facial behavior features, which are stored in temporary local storage through a SQLite database. Upon clicking the 'Stop button,' the database is automatically synced to Google Cloud Storage for further processing and research purposes over a secured connection, and, upon successful sync the database is discarded from the user's device. The participants are provided with a 'Survey button' following stopping the data collection, which, upon clicking it, opens a Google form with study survey instruments. The responses to the survey are stored in Google Forms and Google Sheets.

The synced low level facial feature frames in the database and survey responses are then consumed by other applications, such as Jupyterlab, for analysis, building dashboards, study compliance monitoring, and student modeling through Google Cloud API.

### 3.2 Facial Behavior Feature Modules

Our processing pipeline begins by detecting the user's face on each camera frame. We use a dlib[24] face detector, which returns a face bounding box if a face is detected or, otherwise, drops the frame. These face bounding boxes are then used to crop faces from the original frames, which are later used in processing. The cropped face is then fed to a dlib shape detector to extract 68 facial landmark points, an AU predictor to estimate AU occurrence, and a facial expression recognizer to detect facial expressions. The facial landmarks are used to calculate head pose and Eye-Aspect Ratio (EAR). The rest of this section describes the raw and derived features in greater detail.

*3.2.1 Action Units (AU).* Action units (AUs) are used to describe the movement of individual muscles or groups of muscles by their appearance in the human face, detailed in the Facial Action Coding System [7]. The FacePsy framework implements an existing model architecture and weights from [15], which predicts the probability of any AU occurring in a frame with a face. Our system uses the following Action Units: *AU1*, *AU2*, *AU4*, *AU6*, *AU7*, *AU10*, *AU12*, *AU14*, *AU15*, *AU17*, *AU23*, and *AU24*. The module was trained in the Expanded BP4D+[47] dataset and achieves an average AUC of 0.866 (F1=0.599) across all 12 AUs.

*3.2.2 Face Expression.* Facial expressions have been shown to be a proxy for communicating feelings and cognitive mental states [14]. For facial expression recognition, we implement a CNN network [42] and train it with the FER 2013[ref] dataset to infer a person's discrete facial expression in seven categories: Neutral, Anger, Happy, Surprise, Sad, Disgust, and Fear, with an accuracy of 66.37%.

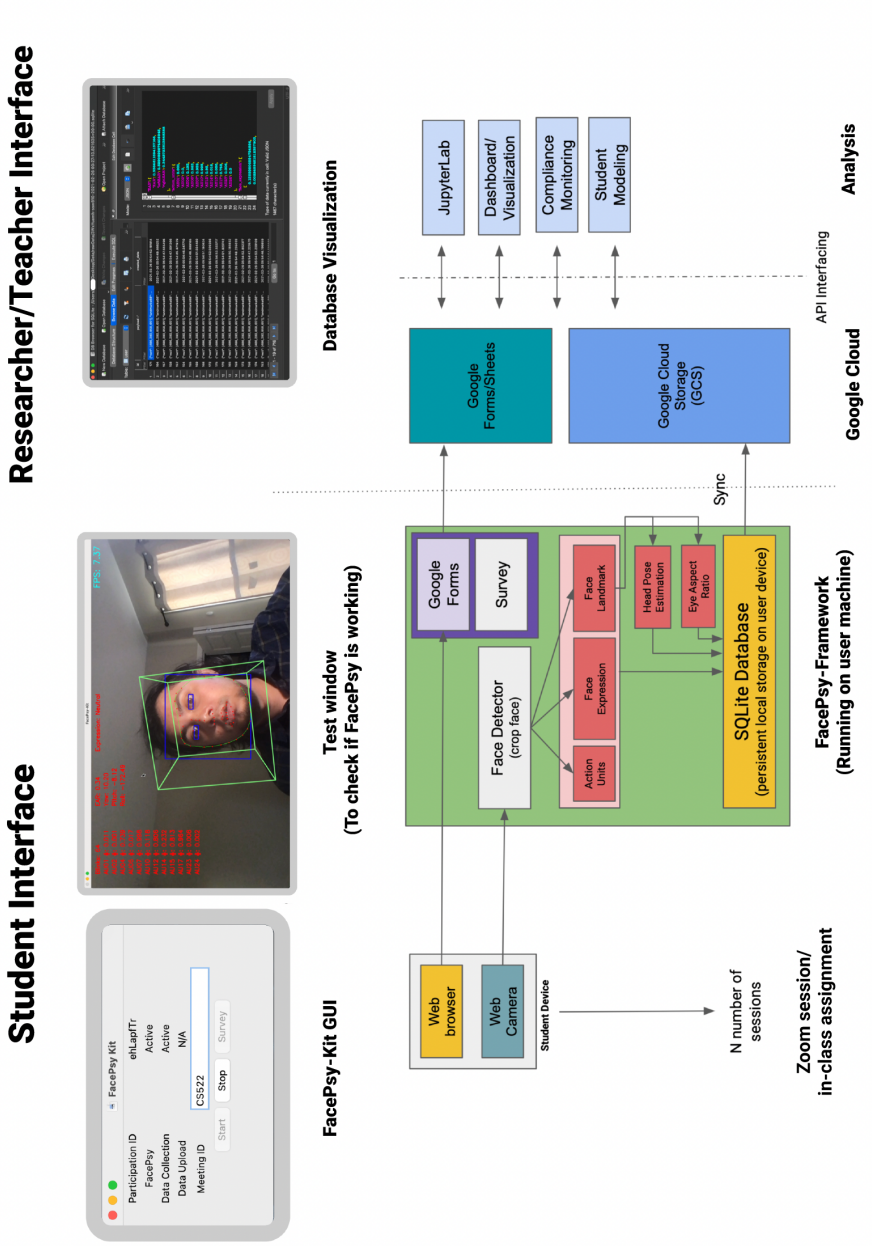


Fig. 2. System Architecture for the FacePsy Framework

*3.2.3 Head Pose.* The head pose is defined in the Euler angles (pitch, yaw, and roll) which parameterize head configurations with respect to the webcam. The yaw angle represents head movement left or right, pitch represents

Table 1. A List of Facial Behavior Features Extracted by Our FacePsy Framework

| Feature Module | Features |
|---|---|
| AU Detector | *AU1* (Inner Brow Raiser), *AU2* (Outer Brow Raiser), *AU4* (Brow Lowerer), *AU6* (Cheek Raiser), *AU7* (Lid Tightener), *AU10* (Upper Lip Raiser), *AU12* (Lip Corner Puller), *AU14* (Dimpler), *AU15* (Lip Corner Depressor), *AU17* (Chin Raiser), *AU23* (Lip Tightener), and *AU24* (Lip Pressor) |
| Facial Expression | Neutral, Anger, Happy, Surprise, Sad, Disgust, and Fear |
| Head Pose | Euler angles i.e. pitch (head nods), yaw (head turns), and roll (lateral head inclinations) |
| Eye-Aspect Ratio (EAR) | Left EAR, Right EAR, and Combined EAR |

head movement up or down, and roll represents head tilts left or right towards the shoulder. The performance of head pose estimation is very sensitive to facial landmark detection. Head movements, such as involuntary head nods, are correlated with state of thinking [14], and head tilt with a state of frustration [32]. Similarly, increased head movements are observed following making a series of errors [44]. Both amplitude of head movement and rotation have a moderately positive correlation with engagement [37].

*3.2.4 EAR(Eye -Aspect Ratio).* EAR is an estimate of the eye-opening state, it is a measure of the aspect ratio of the eye region. Research suggests that eyes are closed for longer periods of time when users experience negative emotions or when they are struggling emotionally [40].

## 3.3 Privacy Preservation

Privacy preservation was one of the primary concerns when creating the system architecture. Given the context of the study, privacy of individual stakeholders is of the upmost importance. The implemented behavior sensing features preserve user privacy by removing any identifiable data from the raw features. The framework uses camera feed to extract lower order facial features, such as facial landmarks, facial expressions, AUs, and head pose, which do not have any identifiable data. These non-identifiable facial features are then transferred to remote server over a secure line. Raw images are then dropped and discarded from user device.

## 3.4 Configurability and Modularity

Unlike other frameworks working in similar domains, FacePsy is incredibly versatile. The framework has an embedded, end-to-end configurable protocol and a large amount of distributed systems working in conjuncture with each other. This high degree of reconfigurability creates room for the creation of unique study protocols. The modularity in the incorporated MLAI algorithm further extends the possibility of creating different sub-frameworks through optimization, fine-tuned to cater to particular domains

## 3.5 Open Source and Community Engagement

FacePsy is an open-source framework developed with the aim to accommodate several egalitarian methodologies. Our core idea is to create a system that is readily employable for data collection and has the potential for use in topics related to classroom engagement, learning management systems, professional development, and knowledge transfers in organizations.

The secondary vision of FacePsy framework is to act as a springboard, facilitating further research and development in the field of intelligent user analysis. Lastly, FacePsy is designed keeping modularity in mind, embarking the community to use, contribute, maintain and improve new features and modules.

# 4 DATASET

## 4.1 Study Deployment

This study was approved by the university's Institutional Review Board. All participants were provided with study guidelines and training before participation. The study was introduced during online classes through Zoom in the beginning of the 2021 Spring semester, and students were required to to provide informed consent to join. The informed consent included an overview of the type of data collected, privacy preservation and confidently policies during data collection, study compliance, and course completion. Although the students were encouraged to participate with an extra credit incentive, participation in the study was voluntary, and students were able to drop out any time. Students were also informed that their participation would not impact their grade or course completion.

As part of the study participation, students were asked to install our data collection application, the FacePsy framework, on their device, and take an online survey at the end of a coding session. At the time of deployment, the FacePsy framework was available for Windows (10) and Mac (Intel CPU), and the participants were required to have a device with a front-facing webcam. After completing their coding task, participants filled out an end-of-session survey, which included Csikszentmihalyi's 3-channel flow Model as an instrument for flow measurement. Table 2 shows our end session survey items. The survey had 3 questions for each item, and participants rated to what degree they agreed with statements either directly (e.g. "I was fully immersed during the coding session") or indirectly (e.g. "The coding task challenges my capabilities to their limits") mentioning a specific item. Final scores for each flow dimension were computed by taking the average rating over each item (n = 3), which were ranked on a scale from 1-10.

Table 2. Students' Self-Report Items based on the Csikszentmihalyi's Flow Model

| Categories | Survey Items |
|---|---|
| Flow | I was fully immersed when programming |
| | I was active during the coding session |
| | The coding task challenges my capabilities to their limits |
| Anxiety | The coding activity makes me anxious |
| | I'd enjoy programming during the activity more if I were more skilled at it |
| | When I encounter a problem in programming I get stuck because I don't know what to do next |
| Boredom | The coding activity is boring |
| | I've lost interest in programming during the session lately because I'm too skilled |
| | Programming during the session isn't as challenging to me as it used to be |

## 4.2 Dataset Description

We deployed the FacePsy framework for data collection in the real-world online learning settings. Over the course of one semester, the FacePsy framework collected facial behavior data from undergraduate and graduate students while they worked on Python coding tasks. In total, we collected 54 end of session survey reports from 19 students, with each session representing a student working on the activity. Students were asked to turn on their FacePsy app for data collection at the beginning of their task and to stop at the end of task, followed by an end session survey. While most of the students complied to the study, completing both the data collection and survey, a few only completed the end session survey. In total we found 19 total surveys without any accompanying sensor data from 3 students. Furthermore, there were 4 sessions for which the duration was less than 5 minutes,

from 4 students, which were dropped. After applying these exclusion criteria we had a total of 31 sessions from 12 students in our dataset. The average number of session across all participants was 2.5 sessions.

To provide an accurate representation of the average session length of participants, we calculated both the mean and median session length for the raw data, which were found to be 73.02 minutes and 31.31 minutes, respectively. Looking over the distribution of session lengths, we noticed that a majority of students (15/31) completed their session in 30 minutes or under, and only 5 of the 31 sessions exceeded 85 minutes. Excluding these 5 sessions (with respective lengths of 180, 153, 150, 166, and 180 minutes) and taking the new average gives us a value of 31.35 minutes. We believe that, because 30 minutes is both the amount of time a majority of sessions were completed in and the closest session to the median and modified average session length, any data beyond this mark was unnecessary. We created a subset from by artificially cutting the original dataset at 30 minutes from the onset of the session. The resulting preprocessing, modeling, and results all occurred using data from this newly modified 30 minute dataset.

In the final dataset, 398,126 student facial feature-frames were sampled every second. Because the FacePsy framework extracts and processes facial images at a rate of 7 FPS and the median session length was approximately 30 minutes, we found it impractical to incorporate all of the raw feature data into the dataset. We decided that one minute was the best unit of analysis, as it has the ability to showcase how a participant's facial features change over the course of a session without an abundance of data to analyze. So, we built and applied an aggregation pipeline, with a frequency of one minute, that computes statistical summaries (sum, minimum, median, maximum, mean, standard deviation, quartile 1, and quartile 3) for each individual participant and session. Each entry of the dataset consists of the statistical summaries of every feature, calculated over one minute. Each entry was then labeled with ground truth surveys by mapping an entire session to its respective survey.

## 5 FEATURE TRENDS AND CORRELATION ANALYSIS

### 5.1 The Selection of Key Features

To examine whether or not a relationship existed between the feature data collected with the FacePsy framework and a student's flow report, we conducted a correlation analysis. To ensure the features selected had a notably strong relationship with flow, we only selected those with a correlation coefficient of .25 or higher. Of the 200 possible features, 14 features had a correlation coefficient greater than or equal to .25. The specific features and their respective correlation coefficients can be seen in Figure 3. Out of these features, 6 were from EAR, 3 were from head pose, 3 were from facial expressions, and 2 were from the AUs.

### 5.2 Features in Different Flow Classifications

After discovering the relationship between these 14 features and reported flow, we wanted to further investigate if they have the ability to differentiate between flow levels. We created comparative box plots that included error bars for each of the 14 features, with one incorporating data from sessions classified as having a 'high' flow level and the other incorporating data from 'low' flow level sessions. A session was described as high flow if the reported flow level fell above the median flow measurement for all sessions (6.67). A low flow session, on the other hand, had a reported flow level less than or equal to the median. To understand whether there was a significant difference between the high and low sessions, we conducted a t-test for independence for each feature. As seen in Figure 4 and Table 3, there was a statistically significant difference for each feature depending on the students' subjective low or high flow reports. Any outliers in the boxplots in Figure 4 were removed solely for the sake of better visualization.

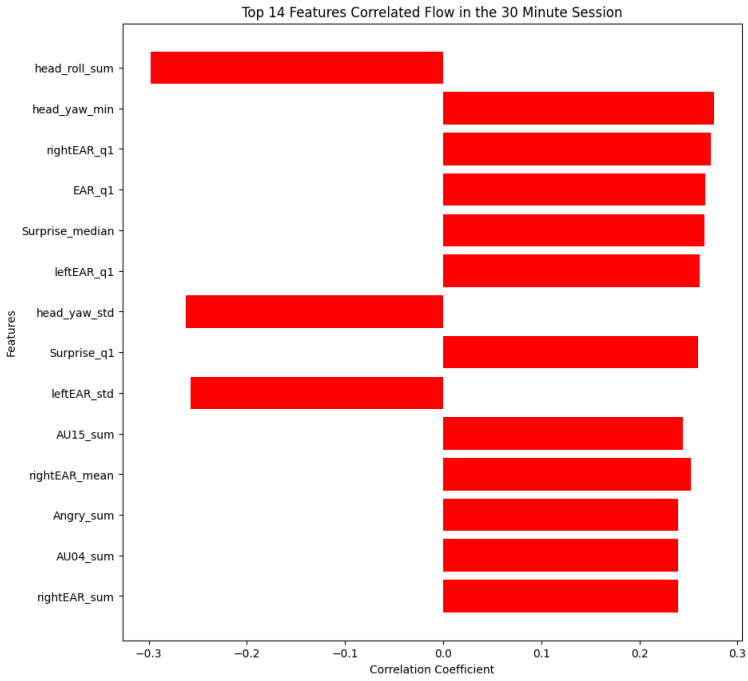


Fig. 3. Top 14 Features Correlated with Students' Subjective Reported States of Flow

## 5.3 Statistical Relationship between Features in Different flow Classifications

This section summarizes the correlations results, as seen in Figure 3, and the t-tests for independence, as seen in Table 3.

*5.3.1 Eye Aspect Ratio.* 6 EAR features were found to be significantly correlated with flow level: right EAR Sum (r=.250), right EAR mean (r=.252), right EAR quartile 1 (r=.273), EAR quartile 1 (r=.267), left EAR quartile 1 (r=.261) and left EAR standard deviation (r=-.2558). Students who reported a high level of flow appeared to have a significantly higher right EAR sum (t=9.92), right EAR mean (t=6.00), right EAR quartile 1 (t=5.73), EAR quartile 1 (t=5.36), and left EAR quartile 1 (t=5.04), and a significantly lower left EAR standard deviation (t =-4.52) than those who reported a low level of flow (p<.001 for each).

*5.3.2 Head Pose.* 3 head pose features were found to be significantly correlated with flow level: Head Yaw standard deviation (r=-.262), Head Roll sum (r=-.298), and Head Yaw minimum (r=.276). Students who reported a high level of flow appeared to have a significantly lower Head Yaw standard deviation (t=-7.32), a significantly higher Head Yaw minimum (t=7.02), and a significantly lower Head Roll sum than those who reported a low level of flow (p<.001 for each).

*5.3.3 Facial Expressions.* 3 facial expression features were found to be significantly correlated with flow level: Angry sum (r=.250), Surprise median (r=.266) and Surprise quartile 1 (r=.260). Students who reported a high level of flow appeared to have a significantly higher Anger sum (t=7.74), a significantly higher Surprise median (t=3.72), and a significantly higher Surprise quartile 1 (t=3.70) than those who reported a low level of flow (p<.001 for each).

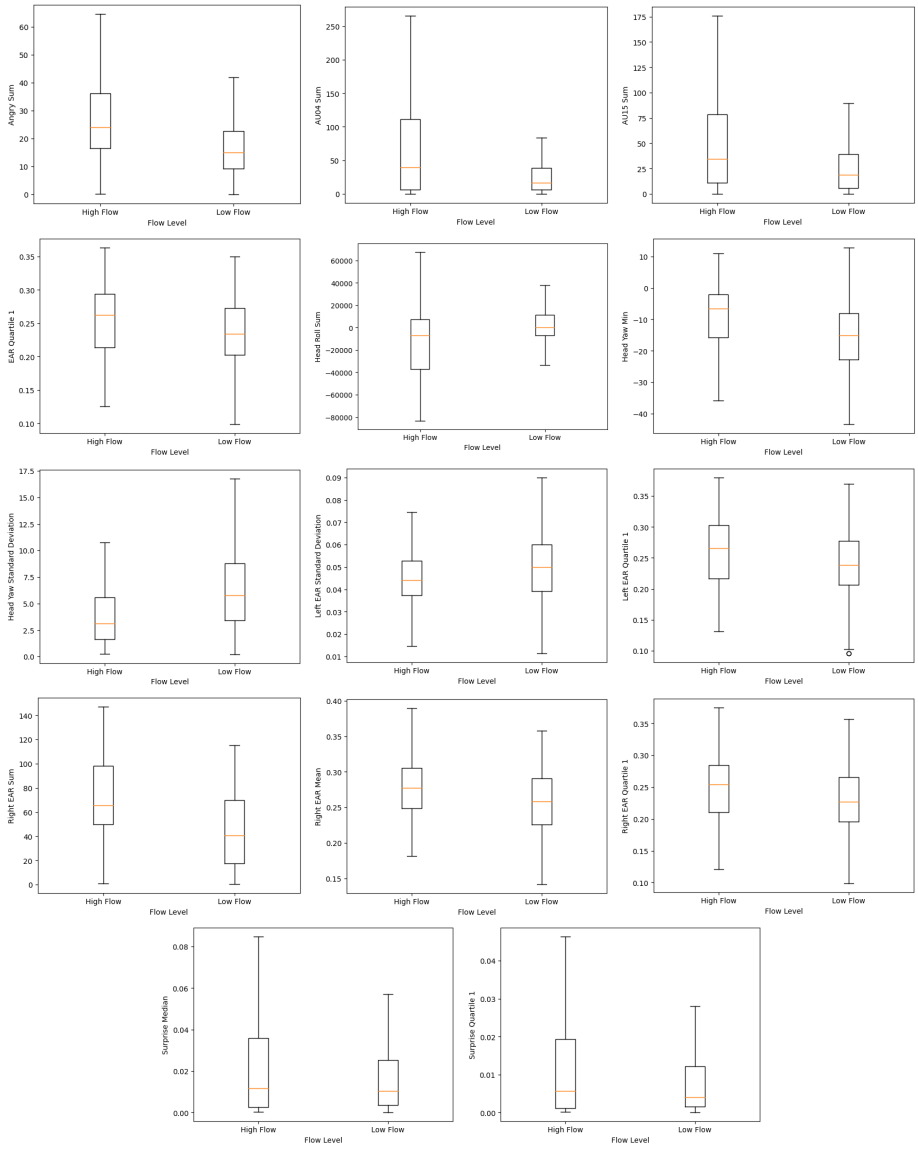


Fig. 4. Comparisons for Each of the Top 14 Features When Students Reported High versus Low Flow

*5.3.4 Action Units.* 2 action unit features were found to be significantly correlated with flow level: AU15 sum (r=.251) and AU04 sum (r=.25). Students who reported a high level of flow appeared to have both a significantly higher AU04 sum (t=6.09, p<.001) and AU15 sum (t=7.29, p<.001) than those who reported a low level of flow.

In summary, we found that an individual in a high state of flow is more likely to have higher EAR, AU04 and AU15, and Anger and Surprise measurements than an individual in a low state of flow. Alternatively, an

Table 3. T-Test Results for High and Low flow Sessions

| No. | Features | High flow | Low flow | t | p |
|---|---|---|---|---|---|
| 1. | Right EAR Sum | M = 72.11<br>SD = 36.97 | M = 45.63<br>SD = 31.10 | 9.92 | <.001 |
| 2. | Angry Sum | M = 27.33<br>SD = 16.22 | M = 18.04<br>SD= 14.36 | 7.74 | <.001 |
| 3. | Head Yaw Standard Deviation | M = 4.03<br>SD = 3.065 | M = 7.38<br>SD = 7.29 | -7.32 | <.001 |
| 4. | AU15 Sum | M = 60.34<br>SD = 71.63 | M = 29.37<br>SD = 33.19 | 7.29 | <.001 |
| 5. | Head Yaw Min | M = -9.51<br>SD = 10.10 | M = -17.42<br>SD = 16.89 | 7.02 | <.001 |
| 6. | Head Roll Sum | M = -13470.14<br>SD = 35607.31 | M = 1162.35<br>SD = 22434.1 | -6.38 | <.001 |
| 7. | AU04 Sum | M = 66.07<br>SD = 72.19 | M = 33.61<br>SD = 47.85 | 6.09 | <.001 |
| 8. | Right EAR Mean | M = .277<br>SD = .041 | M = .257<br>SD = .044 | 6.00 | <.001 |
| 9. | Right EAR Quartile 1 | M = .25<br>SD = .05 | M = .23<br>SD = .049 | 5.73 | <.001 |
| 10. | EAR Quartile 1 | M = .26<br>SD = .05 | M = .23<br>SD = .049 | 5.36 | <.001 |
| 11. | Left EAR Quartile 1 | M = .26<br>SD = .053 | M = .24<br>SD = .051 | 5.04 | <.001 |
| 12. | Left EAR Standard Deviation | M = .045<br>SD = .012 | M = .051<br>SD = .018 | -4.52 | <.001 |
| 13. | Surprise Median | M = .04<br>SD = .085 | M = .022<br>SD = .042 | 3.72 | <.001 |
| 14. | Surprise Quartile 1 | M = .026<br>SD = .062 | M = .011<br>SD = .032 | 3.70 | <.001 |

individual who is in a low state of flow is more likely to have more pronounced Head Pose measurements than those who are in a high state of flow.

## 6 FLOW DETECTION MACHINE LEARNING MODEL

### 6.1 Data processing

The top 14 facial features correlated with flow, whose correlation coefficients were all greater or euqal to .25, were selected to train the model. In order to for our data set to match this selection, all of the columns except for the top 14 features, flow, and Participant ID were removed from the dataset. The data was further cleaned by removing any rows that were missing data. In total, after cleaning there were 652 rows in our final dataset. The MinMax scalar was then used to normalize the data before feeding it to the Machine Learning (ML) model.

Table 4. Light GBM Parameters after Applying Optuna Hyperparameter Tuning

| Parameter name | Search space | Best value | Description |
|---|---|---|---|
| lambda_l1 | (1e-8, 10.0) | 0.0005 | L1 regularization |
| lambda_l2 | (1e-8, 10.0) | 0.1063 | L2 regularization |
| num_leaves | (2, 256) | 217 | Number of branches in the tree |
| feature_fraction | (0.0, 1.0) | 0.8515 | Used to speed and suppress overtraining of the learning |
| min_child_samples | (5, 200) | 54 | Minimum number of data points needed in a child (leaf ) node |
| learning_rate | (0.0, 1.0) | 0.6310 | Learning rate |
| max_depth | (5, 200) | 191 | Limit the depth of the tree structure |

### 6.2 Flow Model Evaluation Over Leave-One-Person Out-Cross-Validation

We further pursued to develop a model that predicts an individual's flow levels during a coding session, preferably as early as possible in that session. To rigorously evaluate our model we decide to use leave-one-person out cross validation (LOPOCV). We took the entire 30 minute dataset and mapped the data from each participant to their flow ground truth values. Then a LightGBM Regressor [23], an implementation of boosted trees, was to trained to predict flow levels for a given session. In Table 4 we present our hyperparameter search space and the best value for each of the parameters. We use the hyperparameter tuning framework, Optuna, [1] to tune our LightGBM regressor with an objective function of minimizing the Root Mean Squared Error (RMSE) and boosting type as Gradient Based One Side Sampling (GOSS). In total we ran 10000 iterations in the hyperparameter search space to find best parameters for our regressor.

To showcase the effectiveness of both the hyperparameter tuning and our feature selection process, we created comparisons between two models tuned in the exact same fashion but trained with the top 14 features and all 200 features, respectively. Each model's accuracy, F1, RMSE, precision, and recall can be seen in Table **??**, with a better visualization in Figure 5.

Overall, we found that the model trained with the top 14 features detected subjective state of flow with an accuracy of 92.65%, which is significantly higher than the accuracy of the model trained with all 200 features (52.66%). Using only the top correlated features improved model F1 by 41%, accuracy by 40%, precision by 48%, recall by 23%, and RMSE by 41%.

## 7 DISCUSSION

### 7.1 Our Discoveries

Using the feature data extracted with the FacePsy framework, we successfully created a machine learning model that can detect a student's flow with 92.66% accuracy. This model trained using the top 14 features correlated with flow from data cut at the 30 minutes from onset mark. This model is successful when compared to the model trained using each of the 200 features, whose accuracy of 52% barely exceeds random chance. Although the error of this successful model has room for improvement, these results suggest that FacePsy can be successfully incorporated into online learning environments for flow detection.

It appears that the facial features extracted by the FacePsy framework can successfully differentiate between a high and low flow level. According to our statistically significant t-tests, an individual experiencing a low level of flow is less likely to present as Angry or Surprised, is less likely to have a lowered brow (AU04) and/or depressed lip corners (AU15), and will have more squinted or closed eyes. They are also more likely to be moving their head, regardless of direction, as opposed to an individual experiencing high flow, whose head will most likely

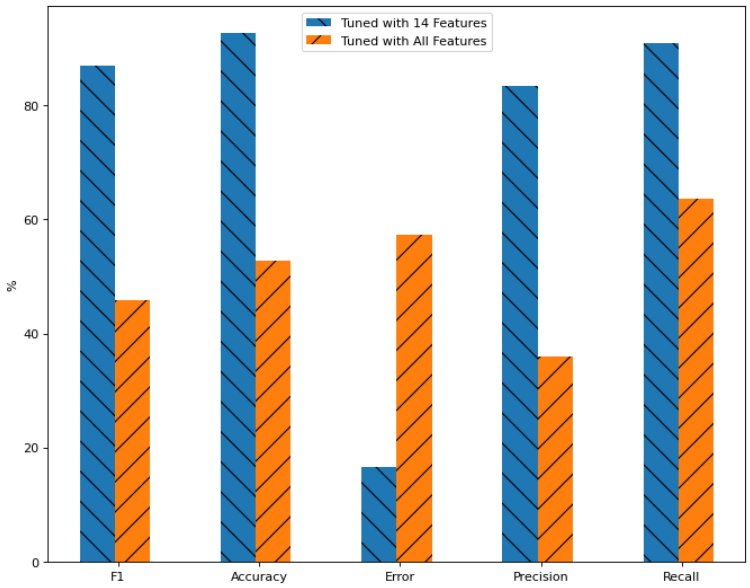


Fig. 5. Model Comparison: by Taking 14 Features versus All Features

stay stationary. Specifically, excess horizontal head movement (yaw) and lateral head movement (pitch) are more likely to be prevalent in an individual with low flow. While flow is considered a 'positive' mental state, it is correlated with traditionally 'negative' emotions. However, the facial muscle movements associated with these emotions in facial recognition methods (lowered brow and lip corners for Anger and widened eyes for Surprise) can be seen in the AU correlations. Because the positive mental state flow is not notably strongly correlated with the traditional 'positive' emotions, facial detection alone may not be sufficient for flow detection.

## 7.2 Relation with Existing Literature

A number of features found in the top 14 correlated features have been seen in similar affective state detection studies. Both Grafsgaard et al. [19] and Booth [4] incorporated AU04 as a feature into their engagement detection methods, although Grafsgaard et al. found that AU04 was correlated with frustration. Krithika found that excess head movement suggested a low level of concentration [25], and Huyunh et al. found that excess head movement suggested a low level of engagement [22], similar to our findings that excess head movement suggests a low level of flow. Although other studies incorporate features similar to ours [2, 36, 45], the specific features used in either model creation or other methods were unclear or not specified. Many of the affective state detection studies most comparable to ours, including those with comparable feature selections, (Aslan et al. [2], Bosch et al. [6], Whitehill et al. [45], and Grafsgaard et al. [19]) extract facial features from recordings of participants, and only Aslan et al. and Bosch et al. use methods that do not require any extra devices. Again, while both of those studies implement similar features, the specific ones used are not directly mentioned.

Interestingly, facial features related to EAR are seldom mentioned in affective state detection. Aslan et al. incorporate eye gaze into their feature selection, but they do not specify if gaze was a feature used in their model [2]. Krithika found that any variation in left or right eye suggested a medium amount of concentration [25]. It is interesting then that almost half of our top features correlated with flow were from the EAR category. Future work that incorporates EAR as a feature in affective state detection may provide further insight into its usability.

This study uses participant self-reports as ground truths to avoid any potential bias that might arise from using observer evaluations. We believe that self-reports will allow for a model that more accurately detect what a student is feeling because it was trained with data from actual students. Furthermore, especially in the online learning context, it is difficult for an observer to accurately gauge another person's emotional state. Observers, as seen in [4, 6, 25, 45] may have inherent biases that affect the ground truths and any subsequent analyses, and some individual's faces may present as angry or frustrated when they are actually engaged or in flow, which is almost impossible for an observer to know.

### 7.3 Potential Impacts

The implementation of the FacePsy framework for flow detection will positively impact both the students who use it and the instructors who implement it. The FacePsy framework's privacy-preserving and non-invasive facial feature detection methods were designed to appeal to students in online learning environments. With this framework, students, who are disinclined to turn on camera in virtual meeting settings, can simply turn on the system without the worry and anxiety that comes with the knowledge of external observers 'judging' them. Instructors can use the data from this framework to better understand how students actually feel about their abilities and progress in a course, which is especially helpful for students who are less likely to actively participate in class. With this better understanding, instructors can tailor lessons and assignments to specifically address problem points and eventually increase a student's learning. These notions have the potential for application in all learning and/or training environments. The FacePsy framework may be able to help evaluate an individual's flow during interviews, presentations, coaching, and more.

## 8 LIMITATIONS AND FUTURE WORK

While our initial performance results are promising, there are some limitations of this study's feasibility that should be mentioned. First, while our dataset provided strong results, it only included 31 unique sessions from 12 individual students. This relatively small size may limit the generalizability of our model. Student participants showed reduced compliance towards the end of the semester. Although an extra credit incentive was provided, the motivation strategies used have room for improvement. Incorporating other incentives to motivate consistent completion will most likely improve both facial behaviors and survey (ground truth) data collection. Thirdly, our FacePsy framework gave students the ability to activate or deactivate the data collection application at any time during their coding session. Students may have clicked the 'Stop' button if they felt that it caused their desktop or laptop to run slower. Students might have also forgotten to initiate the app before starting their coding session. We belive a future deployable system with the ability to automatically control the start and end time of coding activities could improve both the amount and quality of future data collected. Finally, our facial behavior sensing application, although useful, cannot record a student's work space during a coding session. Because we collected data when students participated in the in-class programming activities in their home contexts, it is likely that some participants engaged in other irrelevant online activities instead of completing the given assignment.

Our next step will be increasing the generalizability of our work. We plan to extend this line of research to collect data at large scale. The models we developed in the proof-of-concept study were population models using facial behavior data from all student participants. We are currently in the process of conducting a new study to detect the lack of flow in students wearing face masks in in-person classes. The study compares observer evaluations to student self-reports to see if the data collected with the FacePsy framework can detect affective states that human observers cannot. Ideally, our work would eventually move beyond detection to the prediction of a student's flow state during online classroom activities. Our major goal for online learning is to enable Just-In-Time Intervention support deliveries to students while they are just beginning to lose their flow state or to students who are stuck, and thus increase the effectiveness of online learning as a whole.

## 9 CONCLUSION

We introduced a new facial behavioral sensing framework, FacePsy, which extracts facial features in real-time, without the need to save recordings of participants for later extraction. In this proof-of-concept study, features detected using the FacePsy framework were successfully used to identify college students' subjectively reported state of flow during naturalistic online coding sessions. With the data extracted from students by the framework, we created a machine learning model with a maximum accuracy of 92.65% (F1 = 86.95). This model, trained with the features most correlated with flow, significantly outperforms another model, identical except for the number of features used to train it (all 200 features were used). These initial results provide a good basis for future implementations of the FacePsy framework. Not only can affective state detection now be conducted in a privacy-preserving way, without a large amount of computational power, it can be used as a basis for and improve intervention delivery for online learning.

## REFERENCES


[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *CoRR* abs/1907.10902 (2019). arXiv:1907.10902 http://arxiv.org/abs/1907.10902

[2] Sinem Aslan, Zehra Cataltepe, Itai Diner, Onur Dundar, Asli A Esme, Ron Ferens, Gila Kamhi, Ece Oktay, Canan Soysal, and Murat Yener. 2014. Learner engagement measurement and classification in 1: 1 learning. In *2014 13th International Conference on Machine Learning and Applications*. IEEE, 545–552.

[3] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 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 (2010), 223–241.

[4] Brandon M Booth, Asem M Ali, Shrikanth S Narayanan, Ian Bennett, and Aly A Farag. 2017. Toward active and unobtrusive engagement assessment of distance learners. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 470–476.

[5] Nigel Bosch. 2016. Detecting student engagement: human versus machine. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 317–320.

[6] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces*. 379–388.

[7] Jeffrey F Cohn and Paul Ekman. 2005. Measuring facial action. The new handbook of methods in nonverbal behavior research. *The new handbook of methods in nonverbal behavior research* (2005), 9–64.

[8] Mihaly Csikszentmihalyi. 2000. *Beyond boredom and anxiety*. Jossey-Bass.

[9] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[10] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1–20.

[11] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21.

[12] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2014), 153–170.

[13] Paul Ekman. 2002. Facial action coding system (FACS). *A human face* (2002).

[14] Rana El Kaliouby and Peter Robinson. 2005. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*. Springer, 181–200.

[15] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. 2019. Cross-domain au detection: Domains, learning approaches, and measures. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[16] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3, 2 (1978), 5.

[17] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. 2020. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–26.

[18] AC Graesser, Bethany McDaniel, Patrick Chipman, Amy Witherspoon, Sidney D'Mello, and Barry Gholson. 2006. Detection of emotions during learning with AutoTutor. In *Proceedings of the 28th annual meetings of the cognitive science society*. Citeseer, 285–290.

[19] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.

[20] Joseph F Grafsgaard, Joseph B Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 42–49.

[21] Juho Hamari, David J Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. 2016. Challenging games help students learn. *Computers in Human Behavior* 54, C (2016), 170–179.

[22] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. EngageMon: Multi-modal engagement sensing for mobile games. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–27.

[23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017), 3146–3154.

[24] Davis E. King. [n.d.]. Dlib-ml: A Machine Learning Toolkit. 10 ([n. d.]), 1755–1758.

[25] LB Krithika and Lakshmi Priya GG. 2016. Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. *Procedia Computer Science* 85 (2016), 767–776.

[26] Kwangyoung Lee and Hwajung Hong. [n.d.]. Designing for Self-Tracking of Emotion and Experience with Tangible Modality. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh United Kingdom, 2017-06-10). ACM, 465–475. https://doi.org/10.1145/3064663.3064697

[27] Po-Ming Lee, Sin-Yu Jheng, and Tzu-Chien Hsiao. 2014. Towards automatically detecting whether student is in flow. In *International Conference on Intelligent Tutoring Systems*. Springer, 11–18.

[28] Gwen C Littlewort, Marian S Bartlett, Linda P Salamanca, and Judy Reilly. 2011. Automated measurement of children's facial expressions during problem solving tasks. In *Face and Gesture 2011*. IEEE, 30–35.

[29] Tiina Lynch and Ioana Ghergulescu. 2017. Large scale evaluation of learning flow. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 62–64.

[30] Marco Maier, Chadly Marouane, and Daniel Elsner. 2019. DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2108–2110.

[31] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing* 10, 3 (2017), 325–347.

[32] Judi McCuaig, Mike Pearlstein, and Andrew Judd. 2010. Detecting learner frustration: towards mainstream use cases. In *International Conference on Intelligent Tutoring Systems*. Springer, 21–30.

[33] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.

[34] Lazaros Michailidis, Jesus Lucas Barcias, Fred Charles, Xun He, and Emili Balaguer-Ballester. 2019. Combining Personality and Physiology to Investigate the Flow Experience in Virtual Reality Games. In *International Conference on Human-Computer Interaction*. Springer, 45–52.

[35] Giovanni B Moneta. 2012. On the measurement and conceptualization of flow. In *Advances in flow research*. Springer, 23–50.

[36] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* 8, 1 (2016), 15–28.

[37] Ryota Ooko, Ryo Ishii, and Yukiko I Nakano. 2011. Estimating a user's conversational engagement based on head pose information. In *International Workshop on Intelligent Virtual Agents*. Springer, 262–268.

[38] Olena Pastushenko, Wilk Oliveira, Seiji Isotani, and Tomáš Hruška. 2020. A methodology for multimodal learning analytics and flow experience identification within gamified assignments. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[39] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Michael Thomas Knierim, and Alexander Maedche. 2018. Got flow? Using machine learning on physiological data to classify flow. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[40] Anna Rogalska, Filip Rynkiewicz, Marcin Daszuta, Krzysztof Guzek, and Piotr Napieralski. 2019. Blinking Extraction in Eye gaze System for Stereoscopy Movies. *Open Physics* 17, 1 (2019), 512–518.

[41] Yusuf Can Semerci and Dionysis Goularas. 2021. Evaluation of students' flow state in an e-learning environment through activity and performance using deep learning techniques. *Journal of Educational Computing Research* 59, 5 (2021), 960–987.

[42] Nishank Sharma. 2018. Facial Emotion Recognition on FER2013 Dataset Using a Convolutional Neural Network. https://github.com/gitshanks/fer2013.

[43] Nishank Sharma. 2020. A generator library for concise, unambiguous and URL-safe UUIDs. https://pypi.org/project/shortuuid/.

[44] Mark Tiede, Christine Mooshammer, and Louis Goldstein. 2019. Noggin nodding: Head movement correlates with increased effort in accelerating speech production tasks. *Frontiers in psychology* 10 (2019), 2459.

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

[46] Wenlu Yang, Maria Rifqi, Christophe Marsala, and Andrea Pinna. 2018. Physiological-based emotion detection and recognition in a video game context. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[47] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3438–3446.