FacePsy: An Open-Source Affective Mobile Sensing System – Analyzing Facial Behavior and Head Gesture for Depression Detection in Naturalistic Settings

RAHUL ISLAM and SANG WON BAE*, Charles V. Schaefer, Jr. School of Engineering and Science, Stevens Institute of Technology, USA

Depression, a prevalent and complex mental health issue affecting millions worldwide, presents significant challenges for detection and monitoring. While facial expressions have shown promise in laboratory settings for identifying depression, their potential in real-world applications remains largely unexplored due to the difficulties in developing efficient mobile systems. In this study, we aim to introduce FacePsy, an open-source mobile sensing system designed to capture affective inferences by analyzing sophisticated features and generating real-time data on facial behavior landmarks, eye movements, and head gestures – all within the naturalistic context of smartphone usage with 25 participants. Through rigorous development, testing, and optimization, we identified eye-open states, head gestures, smile expressions, and specific Action Units (2, 6, 7, 12, 15, and 17) as significant indicators of depressive episodes (AUROC=81%). Our regression model predicting PHQ-9 scores achieved moderate accuracy, with a Mean Absolute Error of 3.08. Our findings offer valuable insights and implications for enhancing deployable and usable mobile affective sensing systems, ultimately improving mental health monitoring, prediction, and just-in-time adaptive interventions for researchers and developers in healthcare.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

Additional Key Words and Phrases: Affective computing, Depression, Machine Learning, Mobile computing, System, Empirical study that tells us about people, Application Instrumentation, Field Study

1 INTRODUCTION

Mental health pertains to emotional, psychological, and social well-being, influencing daily thoughts, feelings, and actions. Mental illness is a leading cause of disability, with an estimated 450 million people affected worldwide [71]. It is also a significant predictor of suicide [67]. Mental disorders usually emerge in an individual's early 20s [53], and their untreated presence can negatively impact academic success, productivity, and social relationships [54, 97]. In the context of COVID-19, the need for social distancing has led to the widespread adoption of telehealth services such as telepsychiatry [22, 72]. Unfortunately, many individuals are experiencing mental health issues during the pandemic, with the highest levels of pandemic-era anxiety and depression observed

---

*Corresponding author.

in 2020 across all age groups, which began to decline in early 2021 [96]. While many COVID-19 restrictions have been lifted, approximately 5% of the U.S. adult population, or about 12 million Americans, are living with co-occurring chronic pain and clinically significant symptoms of anxiety and depression [50]. This ongoing situation underscores the importance of reconsidering how to deliver mental health care effectively at the right time. Personalized psychiatric care, which promotes preventive measures and offers tailored interventions, could help meet these needs, though its availability remains limited [1].

A growing body of psychological studies [27, 39, 40, 90] have suggested that depression is characterized by nonverbal signals such as facial muscle movement, and head gesture, which can be detected automatically without the need for clinical intervention. We refer to these as *"facial behavior primitives"*. Research has shown that mental illness, such as depression, leaves recognizable markers in the facial patterns of an individual [81]. Often, these changes manifest in a person's face involuntarily. Creating an automatic system [17, 86, 93] based on these cues can provide an objective and repeatable evaluation and address problems related to cost and time requirements. Despite the valuable insights gained from these studies, it should be noted that they were conducted in controlled lab environments and recorded videos of individuals' faces. Currently, the real-life implementation of such systems is limited due to privacy concerns [10, 17, 56, 87, 94], unrealistic costs [74], and required computational power [15, 32, 69].

While facial actions have shown promise in lab settings for understanding depression, their application in real-world scenarios remains largely unexplored due to challenges in designing efficient, deployable mobile systems. To bridge this gap, we introduce FacePsy, an open-source mobile sensing system capturing facial features, generating real-time data on facial behavior landmarks, eye open, smile, and head gestures, all through smartphones in natural settings, while preserving user privacy. We hypothesize that digital biomarkers extracted from facial cues offer valuable insights into an individual's internal emotional and affective state, thereby enabling algorithms to infer depression. In this field study, we gathered data in real-world contexts to assess how and whether the data collected through our framework could demonstrate the potential for predicting depressive episodes in naturalistic environments.

While there are studies that use mobile sensing to track patterns in social and behavioral data by tracking communications, app usage, and GPS data [15, 69, 99], these mobile sensing-based solutions primarily focus on capturing social and behavioral data but disregarding affective signals, which have been shown to be important indicators of depression. These studies have particular challenges and barriers: (1) Mobile sensing limits modeling usage because it requires extensive computation and post-processing. The burden of collecting sensors 24/7 and battery consumption may lead to low compliance. It might not be usable for stakeholders. (2) Wang et al. [98] tried to capture entire face images in the real-world settings to understand depression but reported that they failed to validate the effectiveness of data due to insufficient frames to build a model (one frame when unlocking the smartphone); (3) Tseng et al. collected and analyzed eye patches in detecting alertness [91] not depression, but partially captured a part of the face only (eye). Most recently, MoodCapture [66] was introduced that captures facial images in natural environments for depression detection. This involves analyzing image attributes such as angle, dominant colors, location, objects, and lighting. The utility of MoodCapture for developers seeking to implement similar studies in different settings may be somewhat limited, as the authors have not made their mobile system, dataset, or machine learning pipeline publicly available. Our study complements MoodCapture's work by exploring the incremental utility of mobile sensing for depression detection and advocating for new avenues to develop mental health assessment tools based on in-the-wild images. Our research advances from MoodCapture in terms of data collection mechanisms, on-device processing, privacy awareness, and the facial attributes collected. While MoodCapture

collects facial data when a user responds to survey questions, our study implements a trigger-based data collection motivated by prior literature [92]. This mechanism activates based on user actions, such as turning the screen on/off, opening/closing apps, etc., to start or stop data collection. For on-device processing, we only send the final detected facial behavior primitives (Action Units (AU), smile, eye open state, head Euler angles, and landmarks) to the research server for further analysis. This helps us ensure user privacy and prevent the leakage of facial images by discarding images from user device after processing. Our work also advocates for privacy-aware data collection. This approach is informed by prior literature on nudging [5], informing users about active data collection [20, 30], and notifying users when the app restarts itself after a reboot or crash [20]. While MoodCapture has first introduced the use of facial images for depression detection in natural environments, our approach proposes a novel facet to this field in several key aspects. We are among the first to develop an open-sourced, privacy-aware, trigger-based affective mobile system that captures facial data from users' smartphones and immediately discards the raw images after extracting essential features in near real-time (within 10 seconds). This method not only builds a predictive model of depressive episodes but also ensures that sensitive facial data is not permanently stored on devices, addressing significant privacy concerns and awareness.

To advance affective computing systems, making it applicable in real-world settings, we synthesize a set of affective signals from face which have been well-validated such as facial muscular activities (AUs) as well as proposed new features beyond simple facial expression algorithms that have been unexplored and invalidated in a user's everyday settings. Research questions as follows: (RQ1) What are the important signals of affective biomarkers on depressive episode detection by differentiating depressive and non-depressive episodes, and how can those key features contribute to the model's performance? and (RQ2) Whether and how can an affective mobile system be efficiently designed, tested, and developed to understand a user's mental health, specifically in predicting depressive episodes, in real-world settings? As a result, we identified specific affective indicators as crucial factors for distinguishing between individuals experiencing depressive and non-depressive episodes. These indicators encompass the eye-open state, head pose, smile expression, and specific Action Units (2, 6, 7, 12, 15, and 17). When these features are combined, they exhibit predictive potential for detecting depression episodes, achieving an AUROC of 67% for universal model, while the hybrid model has an AUROC of 81%. It is worth noting that further enhancements in predictive accuracy can be attained through the accumulation of additional data spanning subsequent weeks. These findings represent a significant stride in bridging the existing disparity between controlled laboratory studies and the practical implementation of depression detection through affective mobile sensing systems in real-world scenarios.

As such, we have developed a deployable and usable open-sourced, lightweight, affective mobile system for the HCI community [1]. Our system integrates state-of-the-art facial biomarkers, which have been validated to understand complex mental states and workloads in controlled lab settings. We have further expanded these features for the context of depression detection. The system automatically extracts these features from a user's smartphone in natural environments. This system has the potential to create new avenues for developing mental health assessment tools and behavior modeling based on in-the-wild images. Moreover, our FacePsy system can be deployed in everyday settings, and it is optimized with a sampling rate of 2.5 FPS without any delays on the user's own phones. Further, we provide insights with the experiments with different subset of facial behavior primitives features in detecting depression in naturalistic environments. As we highlight the impact of each subset of features validated in naturalistic environments, researchers

---

[1]Our system source code is available at: https://github.com/stevenshci/FacePsy

and developers can utilize our mobile system to conduct their studies, configure apps for triggering time and frequency, and build their own computational models.

2 BACKGROUND

In this section, we introduce the literature on facial behavior primitives in developing depression inferences and machine learning models (Section 2.1) and provide prior work on depression detection using mobile sensing technologies in the fields of mobile and affective computing communities (Section 2.2).

2.1 Facial Behavior Primitives in Depression

Research on nonverbal facial behavior often shows that individuals with depression usually exhibit fewer happy facial expressions, reduced expressiveness, and less head movement. The less frequent display of happy facial expressions by depressed patients is a commonly observed finding [14, 37, 79, 85]. Various studies also link depression with decreased general facial expression [35, 79] and head movement [3, 33, 51]. One study found that participants with major depressive disorder (MDD) had a significantly reduced transient pupillary response [61], lending support to the potential of detecting depression through these means. Typical symptoms of depression, such as sorrowful expressions and a lack of affective experience, have been characterized by researchers using facial expressions [27]. However, the prevalence of negative facial expressions in depressed individuals is disputed, with contradictory findings presented in various studies. Some argue that depression is characterized by an increase in negative facial expressions [78, 84], while others suggest that depressed people may actually exhibit more positive facial expressions [35, 79].

To diagnose depression, nonverbal signals have been introduced by researchers in affective computing. For instance, Cohn et al. [17] proposed visual signals as non-verbal behavioral features – manually annotated facial action units (AUs) and active appearance model (AAM) features, which are mathematically derived representations of facial images that capture shape and texture variations. They found that participants with high depression severity displayed fewer associative facial expressions (AU12 – lip corner puller, and AU15 – lip corner depressor) and more non-associative facial expressions (AU14 – dimpler), indicating that these traits helped spot depression. Most recently, Valstar et al. [87, 93] have used facial and auditory features to detect depression in pre-recorded videos. Their finding suggests that AU4 (brow lowerer), AU12 (lip corner puller), AU15 (lip corner depressor), and AU17 (chin raiser) are useful for estimating depression severity, supporting existing evidence. These findings highlight the potential utility of automated facial behavior analysis towards predicting of depression. These initial studies were based on recorded video data from consenting participants in controlled environments. In contrast, deploying such technologies in uncontrolled, everyday settings raises valid privacy concerns. However, our approach mitigates these concerns by processing data directly on the device. By leveraging on-device computation for feature extraction, we significantly reduce the privacy risks associated with transmitting sensitive facial data. This method ensures that personal data does not leave the user's device, aligning with privacy-preserving strategies essential for real-world applications.

Researchers in HCI have explored using mobile camera sensors to capture and analyze user data for mental health assessment. For instance, Rui et al. [98] developed a smartphone app that takes photos of users' faces throughout the day, extracting facial expressions and landmarks. However, this approach had limited success. The correlation between facial expressions, landmarks, and mental health was difficult to establish due to the poor performance of facial expression algorithms, landmark detectors, and image quality. In a separate HCI study, Vincent et al. [92] used pupil information to gauge user alertness as an indicator of mental states. A recent study,

Table 1. Studies on Predicting Depression Using Facial Behavior Primitives

Study	Part.	Study Length	Data & Feature Types	Validation Method	Performance Metrics	Research Environment
Cohn et al. [17]	57	7-week intervals	Audio/video, 17 AUs, AAM	Leave-one-out	Accuracy: 79%	Lab
Valstar et al. [94]	292	One video each	AVEC13, LPQ	5-fold	MAE: 10.88, RMSE: 13.61	Lab
Song et al. [87]	84	One video each	AVEC14, 17 AUs, Pose, Gaze	50/50 split	MAE: 5.95, RMSE: 7.15	Lab
Kong et al. [56]	102	N/A	10 photos, Deep learned	7:2:1 split	Accuracy: 98.23%	Lab
Casado et al. [10]	376	One video each	AVEC13/14, rPPG	N/A	AVEC13: MAE: 6.43, AVEC14: MAE: 6.57	Lab
Wang et al. [98]	37	10 weeks	Photos, Eigenfaces, landmarks	N/A	N/A	In the wild
Nepal et al. [66]	177	90 days	AU, Gaze, Head Pose, Rigidity Parameters, and Eye, 2D & 3D Landmarks	5-fold leave-subject-out	Balanced Acc: 61%	In the wild
Our approach	25	4 weeks	12 AUs, Smile, Eye open, Head Pose, EAR, IVA, 133 landmarks	Leave-One-Person-out	Accuracy: 51%, AUC: 67% MAE: 3.26	In the wild
				Leave-One-Day-out	Accuracy: 69%, AUC: 81% MAE: 3.08	

MoodCapture [66], evaluated depression using images taken automatically by smartphone front-facing cameras during everyday activities. This tool analyzes features in the images such as angles, dominant colors, locations, objects present, and lighting conditions. The study showed that a random forest algorithm trained on facial landmarks can effectively distinguish between depressed and non-depressed individuals, and predict raw PHQ-8 scores. The effectiveness of MoodCapture for developers looking to conduct similar studies in various environments might be constrained because the authors haven't released their mobile system, dataset, or machine learning pipeline. Our study advances from MoodCapture in terms of data collection mechanisms, on-device processing, privacy awareness, and the facial attributes collected. While MoodCapture captures images when participants respond to EMA questions, our protocol relies on opportunistic data collection. We gather data when participants interact with their smartphones at specific triggers (See Section 3.2). This allows us to collect information that more comprehensively represents participants' daily lives and emotional states. By using a broader data collection framework, we can conduct a richer analysis of behavioral patterns and emotional nuances that occur during regular phone usage, not just during specific survey responses. Both MoodCapture and our study are primarily designed for depression detection. However, using smartphone cameras to capture images could also extend to assessing other cognitive states, such as alertness, through pupil imaging. These approaches [66, 92, 98], however, raises privacy concerns due to the transmission and processing of facial images on external servers. Our study complements the findings of these studies by implementing all data processing locally on the user's device; we plan to open source our system to the research community, which can be used for further studies in behavior modeling through affective signals.

2.2 Detecting Depression Using Mobile Sensing

Significant progress has been made in detecting depression with mobile sensing. For instance, Chikersal et al. [15] utilized the AWARE [32], an open-source context instrumentation framework. It tracked behavioral data such as Bluetooth, calls, GPS, microphone, and screen status from smartphones and wearable fitness devices to detect depression in college students. Their method achieved 85.4% accuracy in identifying changes in individuals' depression and 85.7% accuracy in detecting post-semester depression over a semester-long (16 weeks) study. In a different study, Asare et al. [69] also used the AWARE framework to monitor behavioral data. They focused on sleep, physical activity, phone usage, GPS location, and daily mood ratings using the circumplex model of affect (CMA) to detect depression. This approach resulted in 81.43% accuracy. Though sensing

systems have proven effective in various applications, their ability to provide real-time insights and interventions is relatively unexplored. This is mainly due to the resource-intensive requirements of pre-processing, feature engineering, and model development, which demand further scrutiny to fully utilize their capabilities. It is worth noting that the systems examined in previous research did not support near-real-time facial feature extraction. This distinctive feature separates our affective mobile system from others. We summarize the findings of these studies in Table 2.

Table 2. Studies on Predicting Depression Using Mobile Sensing

Study	Sensors Used	Results	Findings
Chikersal et al. [15]	Bluetooth, GPS, Screen, Calls, Sleep	Accuracy: 82.3% F1: 78%	Identifies depressive symptoms using data from smartphones and fitness trackers of college students. The study introduce advanced feature extraction technique.
Opoku et al. [69]	Sleep, Activity, GPS, Phone Usage	Accuracy: 81.43% AUC: 82.31%	Classifies individuals as depressed/non-depressed using mood scores and sensor data. Significant differences found in mood, sleep, activity, phone usage, GPS mobility.
Pedrelli et al. [74]	EDA, Heart Rate, Accelerometer, Sleep, Movements, Temperature	MAE: 3.88 - 4.74	Feasibility of monitoring depression severity with smartphones and wearables, showing moderate to high correlations with clinician-assessed scores.
Farhan et al. [29]	GPS, Physical Activity	F1: 55%	Behavioral data from smartphones predict clinical depression. Combining with PHQ-9 scores enhances accuracy.
Nepal et al. [66]	AU, Gaze, Head Pose, Rigidity Parameters, Eye, 2D & 3D Landmarks	Balanced Acc: 61%	Uses smartphone images to detect depression by analyzing facial expressions and features, demonstrating machine learning potential in mental health assessment.
Islam and Bae [48]	Pupil-Iris Ratio	Accuracy: 76% F1: 64% AUC: 71%	Pupillary response in natural settings varies between morning and evening and can differentiate between depressive and non-depressive states.
Our approach	12 AUs, Smile, Eye open, Head Pose, EAR, IVA, 133 landmarks	Accuracy: 69% F1: 67% AUC: 81% MAE: 3.08	FacePsy detects depressive episodes by collecting facial behaviors and head gestures in real-world settings, achieving high predictive accuracy with key features like eye-open states, smile expressions, and specific Action Units.

Pedrelli et al. [74] have collected data streams from a wearable tracker, Empatica, and smartphone to detect changes in depression severity. The authors leveraged physiological data such as electrodermal activity (EDA), peripheral skin temperature, heart rate, motion from the 3-axis accelerometer, sleep characteristics, social interactions, activity patterns, and the number of apps used, etc. They evaluated their predictive models using two evaluation methods: user-split and time-split. They achieved an mean absolute error (MAE) ranging between 3.88 and 4.74. However, their work's limitation is that participants must wear two E4 Empatica, one in each hand. Such obtrusive approaches lead to decreased compliance of participants and extra cost ($1,690 per device) for researchers and users. In another study [29], they have used GPS and physical activity as sensor features for depression. In another study [48], researchers used pupillometry data as a proxy for psychological state to detect depression. They found pupillary response in natural settings varies between morning and evening and can differentiate between depressive and non-depressive states. All of the earlier studies have mostly focused on behavioral and social changes in a person during the depression because open-sourced and deployable mobile frameworks are limited in the HCI community. As we know, depression is a multifaceted disorder that affects the behavioral, physiological, and social aspects of people's lives. The current mobile sensing-based approaches [15, 29, 69] may not be able to capture the physiological signals with rich emotional signals, which have been shown to be important indicators of depression. Whereas wearable sensing-based physiological sensing solutions are proposed, the cost is very high for such deployment. As motivated by the works [10, 17, 56, 87, 94] in affective computing have shown great potential in capturing physiological signals with rich emotional data in persons with depression in lab settings. However, these studies have been unexplored in real-world deployment where continuous symptom monitoring is essential part of delivering an appropriate real-time intervention. In this paper, we would address this specific issue by developing an affective mobile system that can unobtrusively and opportunistically track

individual facial behavior primitives to give insights into their complex mental state in near-real time by extracting various emotional data motivated by theories in affective computing. As human face serves as a crucial and natural medium for conveying emotional and mental states [26]. While some studies use facial behavior primitives to detect depression [87, 93], attempts to detect depression using passively sensed facial behavior primitives in a naturalistic environment have been unsuccessful [98]. The effectiveness of the existing depression detection models based on facial behavior primitives in natural environments is also unexplored.

Therefore, we aim to design, refine, and develop configurable triggering for data collection, including when unlocking phones and app use, to capture users' facial behavior data in their everyday settings. Our method opportunistically captures users' facial behavior primitives unobtrusively by triggering data collection when users interact with their own smartphones. In addition to detecting depression, we explore the minimum number of days user data required to produce reliable performance. The following sections describe our approach in detail, starting with the design of our passively running mobile affective system.

3 DESIGN OPEN-SOURCED AFFECTIVE MOBILE SYSTEM FOR RESEARCHERS IN HCI COMMUNITY

Our framework design is built upon the HCI theory and affective computing research. This section introduces an overview of designing our affective mobile system, FacePsy (Section 3.1, 3.2, 3.3, 3.4)., and evaluates the feasibility of the system in a pilot study to refine the FacePsy (Section 3.5). FacePsy is open-sourced with several key objectives. Our primary goal is to encourage widespread adoption of the system, enabling users to derive value from the generated data and facilitating engagement with topics related to mental health sensing, particularly in the context of depression. Additionally, we aim to cultivate a community of contributors who can enhance the system by designing FacePsy with modularity as a central principle. The complete source code for FacePsy can be accessed on GitHub: https://github.com/stevenshci/FacePsy.

3.1 Technical Aspects of FacePsy

FacePsy is designed to capture real-time facial behavior primitives as users interact with their mobile devices. Operating with a response time of 2.5 Hz, which was robustly tested across two different devices, the app leverages the front camera to gather facial data during specified triggers opportunistically. This approach enhances data relevance and optimizes energy consumption and privacy. FacePsy integrates advanced modules such as facial landmark detection [41], head pose estimation [41], and facial action unit recognition [28], running these sophisticated processes directly on the device. This on-device processing ensures privacy and increases energy efficiency by eliminating the need for continuous data transmission. In response to the challenges prevalent in the HCI and Ubicomp domains concerning the deployment of everyday facial behavior sensing systems, FacePsy emphasizes: (1) Achieving high performance in facial image capture and feature extraction without compromising the user experience. The tested response time ensures that the app functions effectively in real-time. (2) Prioritizing on-device image processing to safeguard user privacy and improve battery efficiency. (3) Implementing trigger-based data collection to refine model performance, reducing the need for continuous data monitoring and processing. (4) Enhancing system controllability and configurability, which allows researchers to customize data collection parameters according to specific research needs.

The rest of the section describes the facial behavior primitives detection implemented in the FacePsy system to achieve the required research goal in detail.

3.2 Unobtrusive Background Sensing on A User's Phone

FacePsy has introduced functionalities, including configurable app triggers for data sampling. Researchers can now set data collection trigger conditions and sampling rates, with a default set to 10 seconds, informed by studies indicating peak emotional responses within this timeframe [24, 64]. The app supports real-time feature extraction using the smartphone's camera. We quantized a CNN model [28] to integrate into our system; it efficiently extracts facial behavior features, with processing time varying based on the device's capabilities and feature complexity. Processed images are automatically discarded from the user device after 20 seconds to ensure user privacy and manage storage. None of the processed images leave the user's device or are processed outside of user's device. Additionally, our system offers researchers flexibility in defining sampling triggers and rates, enhancing its applicability for diverse research needs.

Upon installation, the mobile app registers as a background service, monitoring user events like phone lock/unlock and application usage (e.g., WhatsApp, Twitter). These events trigger a 10-second data collection session using a photo burst, a duration optimized for battery and computational efficiency based on our feasibility study. FacePsy captures facial markers, such as Action Units [25], and securely syncs this data to a research server during this session. To balance image processing demands and resource consumption, the app records at a rate of 2.5Hz, ensuring seamless phone usage. This decision was informed by a feasibility study involving two users. An alternative for a higher frame rate could involve recording as a media stream and processing it frame-by-frame using a codec.

3.3 Design Rationale Behind Facial Behavior Primitives Selection

Our framework design is based on principles of HCI theory and is informed by extensive research in affective computing. Theoretical Backgrounds on hand-crafted features (e.g., HOG, LBP, etc.) [21, 88] or deep-learned [49, 101] features have adopted to represent each frame or short video segment in lab settings. However, traditional hand-crafted features are not optimal for facial behavior applications as they are not specifically designed for this purpose. Based on previous studies, which suggest that non-verbal visual cues characterize depression, our proposed approach uses facial behavior attributes such as Action Units (AUs), face landmarks, and head pose as frame-wise descriptors. The machine learning kit (ML Kit) [41] is used to automatically detect face landmarks, smile and eye-open probability, and head pose, facilitating data collection. A convolutional neural network (CNN) [28] is adapted to detect the intensities of 12 different AUs, resulting in 151-channel facial behavior time-series data (12 AU, 1 smile probability, 2 eye open probability, 3 head pose, and 133 face landmarks) for each session.

AUs from the facial action coding system (FACS) [25], which taxonomizes human facial movements, specifically AU4, AU12, AU15, and AU17, have been linked to depression severity [36, 87] and mood disorders [43, 55]. Additionally, 133 facial landmarks that localize key facial regions are detected using ML Kit [76]. Features extracted from these landmarks, such as Eye-aspect ratio (EAR) [31] and intervector angles (IVA) [46], have associations with hypervigilance [8], drowsiness [63], and facial expression analysis [46]. Moreover, they've been instrumental in assessing mental fatigue [13]. We also computed probabilities for smile and eye-open states, which have been linked to depression [37] and fatigue [57, 100]. Lastly, our app captured head Euler angles representing head movements. Head pose features, such as slower head movements and specific head orientations, have been identified as indicators of depression [3, 87] and suicidal ideation [23, 60].

The rationale behind selecting these features is rooted in their established associations with mental health indicators, especially depression. By integrating a wide range of facial attributes, our approach aims to provide a holistic and unbiased representation of facial behaviors. Compared

to previously employed hand-crafted and deep-learned features, the facial behavior descriptors provided in this work have several advantages. They are impartial since their values are unrelated to the subjects' identities, which prevents bias based on gender, age, ethnicity, etc., from influencing the results, as suggested by Song et al. [86]. Second, They have a clear, comprehensible meaning, which makes them more interpretable.

3.4 Configurability, Privacy and User Awareness

To provide high configurability for tailoring the application as needed, the FacePsy helps researchers adjust the duration of data collection, which by default is set to 10 seconds upon any trigger for data collection. Furthermore, it allows for distinct data collection durations based on different trigger types, such as app usage, and phone unlocking. Lastly, the system supports the configuration of various app usage triggers to initiate data collection, ensuring a comprehensive and adaptable research tool.

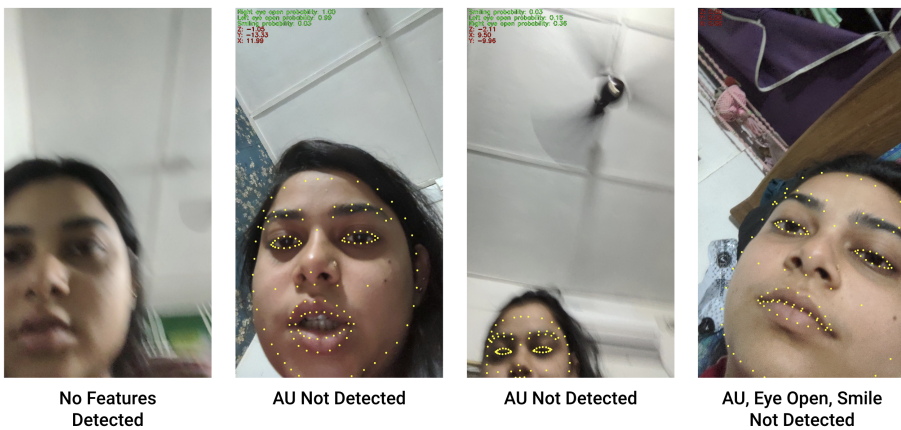| No Features Detected | AU Not Detected | AU Not Detected | AU, Eye Open, Smile Not Detected |

Fig. 1. Examples of Test Images that Fails to Capture Features

When designing our system, paramount importance was given to user privacy and awareness. Drawing inspiration from HCI research, we investigate user nudging [5, 30] in privacy systems [20]. This involves giving information about background data collection, especially in private contexts where no images are stored. Firstly, a clear notification is displayed in the notification bar, indicating "Data collection is active" to ensure users are always aware when data is being collected. Additionally, a green light indicator is positioned next to the camera, signaling when the camera is actively capturing data. In the event FacePsy app automatically restarts itself in the background, either due to a phone restart or an app crash, a toast notification is presented to the user, stating "FacePsy is running on background" Lastly, to further safeguard user data, once facial behavior primitives are extracted, the processed images are automatically deleted from the user's device.

3.5 Feasibility Study

Prior to investigating if facial behavior primitives collected from FacePsy can assess a user's depression status, we decided to carry out a feasibility study to see whether the system could precisely detect the user's facial behaviors using the front-facing camera of a smartphone. To do this, we enlisted the help of 1 volunteer (T2: tester 2) and the first author (T1: tester 1), who used a Google Pixel 4 and 5a smartphone to gather data for two days. Participants carried FacePsy installed in their

| Indoor, Bright light, Seated | Outdoor, Night, Low lighting, Walking | Outdoor, Day, Walking | Indoor, Blurred, Changing position |

Fig. 2. Examples of Test Images That Succeeds to Capture Features

Table 3. FacePsy App Resource Usage on Google 4 & 5a

Resource	Day 1 (T1, T2)	Day 2	Avg (T1 & T2)
Battery	37%, 57%	58%, 43%	48.75%
Memory (MB)	133, 144	85, 57	104.75
Data usage (MB)	9.25, 22.23	16.13, 6.3	13.48
Storage (MB)	175, 155	177, 177	171

phones into their daily lives. FacePsy collected data whenever participants unlocked their phones. We evaluate our system for perceived slowness in their device, causing any delay, interruption, or disruption to phone usage. Our facial behavior sensing modules generated pictures annotated action units, smile probability, eye open probability, head pose, and face landmarks. After gathering these pictures, the first author confirmed if the facial behavior primitives were correctly detected. To do this, the first author manually verified that the annotated photographs matched the unannotated photos. Annotated images of facial behavior markers that were identified are shown in Figures 2, while images were accurate, the processing modules made errors on several occasions (See Figure 1). On average, the app consumed 48.75% battery while running in the background and 104.75mb memory. See Table 3 for more details. In total, we collected 834 images. Our features extraction module processed only 817 images where a face was detected with an average failure rate of 2.04% (See Table 4). This was calculated based on daily end-of-day reports from volunteers regarding resource consumption by our app. To optimize resource consumption, we implement opportunistic data collection, which allows the app to collect data less frequently, thus reducing the load on device resources. Further refining data collection triggers can be done to ensure that the app only collects data when optimal behavioral signals are present, reducing unnecessary resource usage.

The results of our app feasibility evaluation were encouraging. Our system detected the facial behavior makers without hindering the user device experience at a 10-second photo burst in the background.

Table 4. FacePsy Data Processing Performance

Metric	T1	T2
Total Images	411	423
Successful Extractions	401	416
Number of Failures	10	7
Overall Failure Rate (%)	2.43%	1.65%
Failure Rate for AU (%)	10.97%	24.28%
Failure Rate for Landmarks (%)	0.00%	0.00%
Failure Rate for Classification(Eye, Smile)	0.00%	1.68%
Failure Rate for Head Pose (%)	0.00%	0.00%

4 FIELD STUDY DATA COLLECTION

In this section, we describe our study protocol and data collection process in naturalistic environments.

4.1 Participants

Of N=25 participants (mean age 27.88 ± 8.87, range 18 - 48)[2], 8 were females, 11 were males, and 6 did not specify gender on the demographic survey. 15 participants were Asian, 4 participants were Caucasian, and 6 participants did not specify their ethnicity on the demographic survey. 1 participant indicated that their highest education was high school, 8 participants had bachelor's degrees, 10 participants had master's degrees, and 6 participants did not indicate their highest education on the demographic survey. 4 participants indicated that they had been diagnosed with a mental disorder in the past, 15 indicated they had not, and 6 did not answer the question on the demographic survey. Detailed demographic distribution is provided in Table 5.

4.2 Participants and Study Procedure

This study was reviewed and approved by the Institutional Review Board (IRB) at the University. Participants in this study were on-boarded remotely across multiple time zones via Zoom Conference Meetings. Participants were eligible to participate in the study if they were above 18 and owned a data plan-enabled Android smartphone. The research team advertised the study through flyers and posts on Facebook and Whatsapp groups. Participants were asked to respond to a screening questionnaire and select a preferred time for the onboarding Zoom meeting. In the onboarding meeting, the interviewer gave participants informed consent and asked them to respond to the baseline questionnaire. After the baseline, the interviewer took a semi-structured interview to understand the participant's mental health, followed by installing a mobile application on the participant's device to track sensor data from their smartphones. The study questionnaires were delivered through email and administered with Qualtrics, an online survey platform.

Out of 38 participants who were initially recruited, only 25 participants completed the study. One participant reported having a high battery drain because of our app and dropped out of the study after two days. We later found the participant had high social media usage, which resulted in frequent data collection triggers. Among others, 3 participants dropped out for personal reasons, 5 didn't complete surveys, and 4 had incompatible Android versions, leading to the failure of the data collection trigger module. The participants were compensated up to $135 for full compliance with the study. The participants were compensated $20 for baseline and installing the data collection app and were compensated $25 weekly for 4 weeks.

---

[2]9 participants did not provide an age in the demographic survey.

Table 5. Demographic Distribution

Attribute	Unspecified	Male	Female	Total
Gender	6	11	8	25
Age (Average)	-	24.11	32.71	27.88
Ethnicity				
- Unspecified	6	0	0	6
- Asian	0	9	6	15
- Caucasian	0	2	2	4
Education				
- Unspecified	6	0	0	6
- High School	0	0	1	1
- Bachelor's Degree	0	5	3	8
- Master's Degree	0	6	4	10
Mental Health Rate (Average on a scale of 1-10)	-	6.73	6.88	6.79
Depression State				
- Unspecified	6	0	0	6
- Not at all often	0	2	1	3
- Not so often	0	4	4	8
- Somewhat often	0	5	1	6
- Very often	0	0	2	2
Mental Disorder Diagnosis				
- Unspecified	6	0	0	6
- No	0	9	6	15
- Yes	0	2	2	4
Smoking Marijuana				
- Unspecified	6	0	0	6
- No	0	9	8	17
- Yes	0	2	0	2

4.3 Ground-Truth: Mental health measures

Participants' depression symptoms were assessed using a self-reported 9-item Patient Health Questionnaire-9 (PHQ-9) [58] at three distinct times: upon joining the study (baseline), two weeks into the study (mid-point), and at the conclusion of the study (end-point). Each item on the PHQ-9 is scored from 0 (not at all) to 3 (nearly every day), with the total scores ranging from 0 to 27. This scoring captures the frequency of symptoms such as mood, sleep issues, fatigue, and changes in appetite over the past two weeks. Monitoring a single person's multiple PHQ-9 scores over time can yield valuable insights into their mental health progression. The PHQ-9 categorizes depression severity into five levels: scores of 0–4 signify no depression symptoms, 5–9 indicate mild depressive symptoms, 10–14 represent moderate depressive symptoms, 15–19 signify moderately severe depressive symptoms, and 20–27 denote severe depressive symptoms. This method allows researchers and clinicians to track the severity and changes in depressive symptoms effectively.

We can define a depressive episode as a period characterized by persistent feelings of sadness, hopelessness, and a lack of interest or pleasure in most activities observed with a PHQ-9 score during a two-week observation period. we label two weeks of data from the participant as depressed or non-depressed (i.e., a depressive episode) based on the PHQ-9 score of the participant at the beginning and end of the two-week observation period, which falls within the range of mild depressive symptoms or worse according to the PHQ-9 severity scale [38]. We label an individual

FacePsy: An Open-Source Affective Mobile Sensing System – Analyzing Facial Behavior and Head Gesture for Depression Detection in Naturalistic Settings

MobileHCI '24, Sept 30 – Oct 03, 2024, Melbourne, Australia

as having a depressive episode only if the PHQ-9 score of the person is equal to or greater than 5 at both the beginning and end of the observation period; otherwise, it is considered a non-depressive period. This approach ensures consistency in labeling periods as depressive or non-depressive based on established thresholds of depressive symptom severity. We complement our binary classification models by incorporating regression models designed to predict the PHQ-9 scores. It's important to note that the PHQ is a versatile instrument, used both for screening for depression and for monitoring changes in clinical symptoms [59]. In total, we label 14 cases of depressive episodes and 30 non-depressive episodes (where depressive episode length is two weeks). We excluded the last two weeks of data from 6 participants due to non-compliance, resulting in the exclusion of 6 depressive episodes.

4.4 Facial behavior data collection

The FacePsy app activates to capture facial data in three circumstances: when the user unlocks their phone when the user accesses one of a preset number of trigger apps. The phone unlock trigger activates the FacePsy app for 10 seconds after a user unlocks their phone. FacePsy also activates for 10 seconds when the user opens one of thirty-five different apps in total, divided into the categories of communication, social, productivity, entertainment, and health. For example, Instagram, Google Chrome, and Android Messages are all considered trigger apps. The images processed were mostly clear and had high resolution. However, some images had some noise or blurriness on which the model could not detect faces, which is a precursor for routines such as AU detection, landmark detection, etc. These frames were dropped.

5 DATA PROCESSING AND ANALYSIS

5.1 Feature Engineering

Since most of the features are extracted by our behavior-sensing system on the user phone itself, we extract very few features as part of post-processing. Our system pre-extracts features such as Action Units (AU1, 2, 4, 6-7, 10, 12, 14, 17, 23-24), Smiling and Eye open probability, face landmarks, and Head pose features (yaw, pitch and roll) on user device itself. We additionally extract features such as the Eye-aspect ratio and Inter-vector angles. More details on these features are described below.

Inter-vector angle. Inter-Vector Angles (or IVA) are scale-invariant geometric features computed on facial landmarks for the purpose of facial shape representation [46]. We consider the nose center as the centroid of the face for the purposes of computing IVA features. We then segment the face into 8 regions (nose center, jawline, left eyebrow, left eye, right eyebrow, right eyebrow, mouth, and cheeks) and compute in total 1439 triangles by taking permutations of all possible triangles from the centroid to the remaining facial landmarks. We then use Principle Component Analysis to reduce the number of IVA features down to 10. We then compute angular velocity and acceleration.

Eye-aspect ratio. Eye Aspect Ratio is a measure of the aspect ratio of the eye region, which is used as an estimate of the eye-opening state. We defined EAR as the sum of two vertical lengths of the eye divided by two times the horizontal length of the eye.

We collected 12 Action Units (AU), 1 Smile Probability, 2 Eye Open Probabilities, 3 Head Euler Angles, 2 Eye Aspect Ratios, and 20 Inter-Vector Angles. We segmented each participant's day into four epochs: midnight (12am-6am), morning (6am-12pm), afternoon (12pm-6pm), and evening (6pm-12am), each lasting six hours. For each epoch, we computed statistical features such as min, max, mean, median, sum, std, q1 and q3 to summarize the features of that epoch. After this procedure, we have 320 features for each epoch in our final dataset. We then classified each instance into its
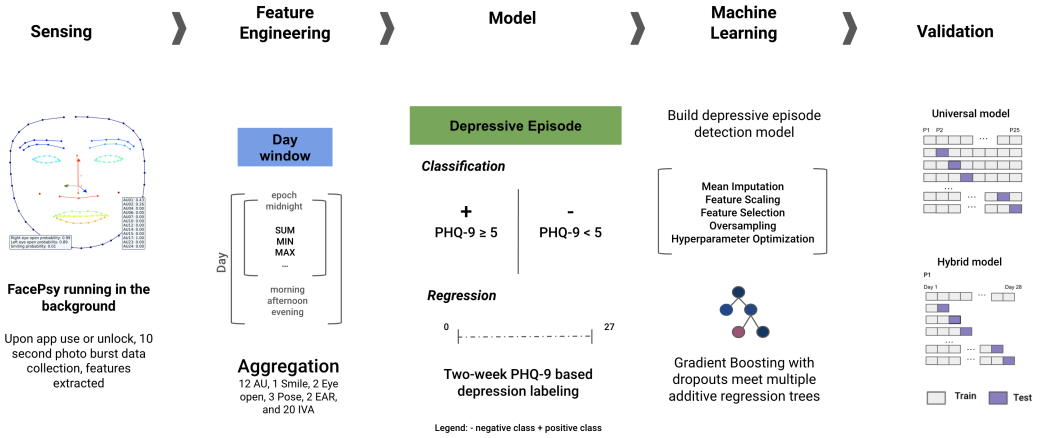
Fig. 3. Overview of Our Affective Mobile System

respective depression class. We noticed variations in participation duration, with an average of 3.36 days missing due to participants' early exit from the study. This resulted in a total dataset of 616 participant days. In total, we gathered 544 days of facial data from 25 participants over a four-week period. Notably, there were 55 days without any recorded data, as explained by participants, such as planned holidays or breaks. A further 17 days lacked data due to issues like image quality and eye-open probability in the feature extraction process. As a result, the total number of effective data points for analysis is 544.

5.2 Statistical Analysis

The primary target variable of interest in our statistical analysis was the presence or absence of a depressive episode. We calculated Pearson's correlation coefficient (r-value) for each feature within our dataset to assess its relationship with the target variable. Additionally, we determined the mean and standard deviation for each feature, separating the data by group. Features were then ordered according to the absolute value of their r-values to pinpoint those with at least a weak correlation (r-value >= |0.20|). This approach allowed us to focus on the most significant relationships, improving the interpretability and efficiency of our models, minimizing unnecessary complexity, and reducing the likelihood of overfitting.

5.3 Feature Selection

Our analysis used feature selection (FS) with a Decision Tree classifier, specifically the CART (Classification and Regression Trees) algorithm implemented in scikit-learn's DecisionTreeClassifier. To compute the importance scores for all features in the dataset, we used the Gini importance, a measure derived directly from the Decision Tree itself. We then established a threshold for feature selection based on the mean value of these importance scores, which calculated to 0.00078125. Features with importance values above this threshold were considered significant and retained for further analysis, while those below were discarded. This approach ensures a data-driven, objective criterion for feature selection, enhancing model interpretiveness and efficiency by focusing on features that contribute to the prediction of depressive episodes.

5.4 Predictive Modeling with Machine Learning

In our predictive analysis using machine learning (See Figure 3), we developed a classification model to detect instances of depression and non-depression, as well as a regression model to predict the PHQ-9 score (See Section 4.3). We assigned labels of depression and non-depression to each day of data based on their corresponding PHQ-9 scores, which served as the ground-truth values.

Our predictive modeling framework utilizes LightGBM (LGBM), a machine-learning library that implements a gradient-boosting algorithm. This method has demonstrated robust predictive capabilities in previous research [69, 70] focused on depression prediction. Our goal is to develop both a universal model that identifies general patterns and a hybrid model that captures intricate interactions and temporal sequences within the data. To address the issue of class imbalance, where depressive instances are less frequent, we applied SMOTE [11] on the training dataset to enhance the representation of the minority class, used mean imputation for handling missing data, and performed standard scaling on the features. Hyperparameter tuning was conducted tomaximize the AUROC value for classification and MAE for regression. We constructed nine supervised models, each tailored to different learning schemes and utilizing subsets of facial behavior features sourced from various facial regions to assess their predictive power for depression. The evaluation of these models primarily relies on the AUROC score [44]. This metric is particularly effective for depression prediction as it evaluates a model's ability to distinguish between depressive and non-depressive states by considering both the true positive rate (TPR) and false positive rate (FPR). The insensitivity of AUROC to class imbalance makes it especially valuable, and a higher AUROC score signifies superior model performance.

5.4.1 Universal model. This learning scheme utilizes a standardized procedure where a single model is created for all users to identify depressive episodes. It uses a leave-one-participant-out (LOPO), a.k.a leave-one-out/leave-one-group-out cross-validation technique. This approach, commonly used in numerous mobile inference systems, provides a clear understanding of model generalizability. Once this universal model is in place, it remains unchanged.

5.4.2 Hybrid model. The ideal model would blend the high precision of individualized models with the ease of use of universal models that don't require user training. Our study was unable to use individualized models due to a lack of data - only two labels per participant, which is not enough for such intricate modeling. We experimented with a hybrid model, incorporating a small quantity of user-specific data with a broader general user dataset. we implement nested cross-validation to minimize the likelihood of model overfitting by implementing a robust ML model training strategy recommended by Asare et al. [69] for the predictive analysis of depression. We employed stratified three-fold cross-validation with a time-series aware leave-one-participant-day-out (LOPDO) cross-validation for the outer and inner cross-validation. In other words, one participant's day is chosen as the test set, and the remaining participant's dataset is chosen as the training set for each iteration of the nested cross-validation. All training set samples captured after the test set are subtracted for time-series awareness. This approach could avoid the unworkable situation in which future datasets are used to forecast the past. Consequently, the LOPDO model effectively combines unique aspects of how an individual's facial behavior data is linked to their state of depression while also identifying general patterns consistently seen across different people. The classifiers' hyperparameter optimization, feature scaling, oversampling, and missing data imputation were all addressed by inner cross-validation. Using grid search across a predetermined set of parameters, we optimized the classifiers' hyperparameters by maximizing the model AUROC score.

6 FIELD STUDY: UNDERSTANDING AND DETECTING DEPRESSIVE EPISODES

To obtain the feasibility of our proposed FacePsy framework in the field study, we address the following question: (RQ1) can the facial behavior features collected by our mobile sensing system be effectively utilized to detect depressive episodes in a naturalistic environment? (Section 6.2). To understand sets of depression-related biomarkers that could be used in the wild we address (RQ2) What's the significance of different biomarkers on depression detection differentiating depressive vs. non-depressive episodes in real-world settings? We introduce the top 27 facial behavior features.

Table 6. Summary of Feature Correlations with Depressive Episodes

Feature	p-value (<0.05)	r-value	Depressive Episode Mean (SD)	Non-Depressive Episode Mean (SD)
ear_right_sum_morning	0.00	0.35	23.34 (28.87)	8.26 (11.1)
ear_left_sum_morning	0.00	0.34	21.36 (25.82)	7.99 (10.83)
headEulerAngle_Y_sum_morning	0.00	-0.33	-404.03 (719.02)	-33.33 (332.65)
leftEyeOpenProbability_sum_morning	0.00	0.33	55.41 (69.64)	21.04 (28.1)
rightEyeOpenProbability_sum_morning	0.00	0.31	48.74 (62.6)	19.99 (25.73)
AU15_min_afternoon	0.00	0.27	0.09 (0.14)	0.04 (0.04)
rightEyeOpenProbability_std_evening	0.00	0.26	0.24 (0.07)	0.2 (0.07)
smilingProbability_sum_morning	0.00	0.26	5.52 (9.62)	1.98 (3.52)
rightEyeOpenProbability_std_afternoon	0.00	0.23	0.24 (0.07)	0.2 (0.07)
AU17_sum_midnight	0.00	-0.23	6.36 (13.6)	21.02 (34.59)
AU12_median_morning	0.00	-0.22	0.29 (0.21)	0.4 (0.27)
AU07_std_evening	0.00	0.22	0.26 (0.08)	0.22 (0.07)
AU02_std_evening	0.00	0.21	0.21 (0.07)	0.17 (0.08)
smilingProbability_max_evening	0.00	0.21	0.35 (0.2)	0.25 (0.21)
AU07_median_morning	0.00	-0.21	0.59 (0.24)	0.68 (0.2)
smilingProbability_max_morning	0.00	0.21	0.33 (0.2)	0.24 (0.2)
smilingProbability_mean_midnight	0.00	0.21	0.14 (0.12)	0.09 (0.09)
AU12_q3_morning	0.00	-0.21	0.39 (0.23)	0.5 (0.27)
AU12_mean_morning	0.00	-0.21	0.31 (0.19)	0.41 (0.24)
leftEyeOpenProbability_std_evening	0.00	0.20	0.23 (0.07)	0.2 (0.07)
headEulerAngle_X_std_evening	0.00	0.20	3.86 (1.4)	3.25 (1.33)
AU06_max_morning	0.01	-0.20	0.56 (0.27)	0.66 (0.23)
smilingProbability_std_evening	0.00	0.20	0.09 (0.05)	0.06 (0.06)
smilingProbability_std_afternoon	0.00	0.20	0.08 (0.06)	0.06 (0.06)
smilingProbability_mean_afternoon	0.00	0.20	0.11 (0.09)	0.07 (0.09)
AU06_mean_morning	0.01	-0.20	0.29 (0.19)	0.38 (0.21)
smilingProbability_mean_evening	0.00	0.20	0.1 (0.08)	0.07 (0.07)

6.1 Statistical difference between depressive and non-depressive episodes

The analysis evaluated the correlation of various features with depressive episodes (See Table 6). The features were ranked based on the strength of their correlation (r-value) with the target variable, which indicates the presence of a depressive episode. Out of 1280 only 158 features had a p-value less than 0.05. Selecting features with at least a weak correlation (r-value >= abs(0.20)), streamlining analysis by prioritizing meaningful relationships, and enhancing model interpretability and efficiency while reducing noise and the risk of overfitting. Temporal dynamics of depressive episodes suggest features measured in the morning often show significant correlations, pointing to the

potential impact of depression on morning routines or states, such as reduced facial expressiveness or specific eye movement patterns. The feature *headEulerAngle_Y_sum_morning* has the strongest negative correlation with depressive episodes, with an r-value of -0.33. This suggests that as the value of this feature decreases, the likelihood of a depressive episode increases. The Y rotation of the head translates to the yaw of the Euler angle. Other features with notable negative correlations include *AU17_sum_midnight*, *AU12_median_morning*, *AU12_q3_morning AU07_median_morning*, *AU06_max_morning* and *AU06_mean_morning*. The absence of *AU06*, associated with expressing emotions related to happiness or joy, correlates with depression. Furthermore, *AU07*, *AU12* and *AU17* have been linked to depression severity, supporting existing evidence [36, 87].

The feature *ear_right_sum_morning* and *ear_left_sum_morning* shows a strong positive correlation with depressive episodes, with an r-value of 0.35 and 0.34, respectively. This indicates that as the value of this feature increases, the likelihood of a depressive episode also increases. It's very important to consider the temporal dynamics of these features. Other features with significant positive correlations include *leftEyeOpenProbability_sum_morning*, *rightEyeOpenProbability_sum_morning*, *rightEyeOpenProbability_std_evening* and *leftEyeOpenProbability_std_evening*. The presence of strong EAR, eye open probability related to high alertness [2] in morning and evening could be explained as the "eveningness–morningness" dimension in depression [12]. The preference for morning or evening can largely be attributed to the reduction of depressive symptoms such as low energy, avoidance of social interaction, and loss of interest in previously pleasurable activities [77].

The presence of a positive correlation (r-values of 0.26, 0.21 for morning sum and maximum, respectively; and similarly positive correlations for evening and midnight measures) between *smilingProbability* and depressive episodes suggests that higher smiling probabilities are associated with an increased likelihood of depressive episodes. This interpretation may seem counterintuitive since it's expected that depressive episodes would be associated with less smiling. However, this unexpected positive correlation doesn't necessarily imply that smiling more leads to depression or vice versa. It might reflect complex underlying behaviors or compensatory mechanisms, such as "smiling depression," where individuals might smile or maintain a facade of happiness in social situations despite experiencing depressive symptoms internally [95]. However, as a limitation of our study we are not able to confirm if participants are going such cases. Previous research suggests that depression is not only associated with sad facial expressions but also with "a total lack of facial expression corresponding to the lack of affective experience" [27]. Since we collect short segments (10 sec) of data, it can be interpreted in various ways, e.g., a smile may be a result of feeling happy or feeling helpless, as suggested by prior research [87]. While a positive correlation between *smilingProbability* and depressive episodes seems paradoxical, it highlights the complexity of depressive behaviors and the importance of considering broader psychological and situational contexts when interpreting these findings.

6.2 Model development from data in the field study

6.2.1 Universal model. Table 7 summarizes the predictive performance of universal models. To understand how different subsets of facial behavior features contribute to detecting depression, we evaluated nine different models, each with a different face feature set, using LightGBM. The model using the most significant features from the correlation analysis performed the best, followed by the model that included feature selection. This approach enhances the model's interpretability and comprehension. The TSF model achieved 51% accuracy, with a precision of 40%, indicating that it correctly predicted depression 40% of the time. A recall of 96% suggests that out of all the depressive episode cases in the dataset, the model successfully identifies 96% of them as positive. This is particularly important in depression detection, where missing out on positive cases leads to missing out on opportunities to intervene. The model's reliability is also reflected in an AUROC

score of 0.67 (Fig. 4). In terms of regression metrics, the MAE for each model also provides insights into the quantitative accuracy of depression severity estimation, showing the lowest error (3.26) for the Action Units model, which suggests its superior ability to estimate the severity of depression correctly compared to other models, where TSF model got an MAE of 5.13. While this indicates limited ability to distinguish between depression and no-depression classes, it represents better agreement between the model's predictions and actual observations than a random classifier.
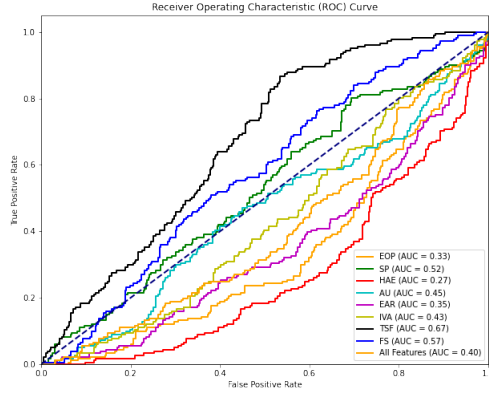

Fig. 4. The ROC plots show the universal model performance of each feature type model.

Table 7. Universal Model Performance: We trained eight LGBM models for predicting depression, including a different feature subset. The model trained using all features showed the best results in predicting depression

Model	MAE	Accuracy	Precision	Recall	F1	AUROC	No. of Features
Eye Open Probability (EOP)	5.20	0.33	0.31	0.83	0.45	0.33	64
Smiling Probability (SP)	5.32	0.37	0.33	0.87	0.48	0.52	32
Head Euler Angle (HEA)	4.71	0.29	0.28	0.71	0.40	0.27	96
Action Units (AU)	**3.26**	0.38	0.30	0.66	0.42	0.45	384
Eye-aspect ratio (EAR)	5.31	0.32	0.30	0.80	0.44	0.35	64
Inter-vector angle (IVA)	4.59	0.40	0.31	0.66	0.42	0.43	640
Top Significant Features (TSF)	5.13	**0.51**	**0.40**	**0.96**	**0.56**	**0.67**	27
Feature Selection (FS)	4.04	0.50	0.38	0.77	0.51	0.57	46
All features	3.77	0.40	0.28	0.51	0.36	0.40	1280

6.2.2 Hybrid model. Table 8 summarizes the predictive performance of hybrid models. The model with the best performance is the one using feature selection with LightGBM. In the context of detecting depressive episodes, the model demonstrated a commendable performance with an accuracy of 69%. Notably, when predicting a depressive episode, it was correct 57% of the time, as indicated by a precision for the depressive class. Furthermore, it successfully identified 62% of all actual depressive episodes, reflected by a recall. The F1-score, a measure of the model's balance between precision and recall, was 0.67 for depressive episodes, suggesting a harmonized performance despite the inherent class imbalance. The model's reliability was also underscored by

FacePsy: An Open-Source Affective Mobile Sensing System – Analyzing Facial Behavior and Head Gesture for Depression Detection in Naturalistic Settings

MobileHCI '24, Sept 30 – Oct 03, 2024, Melbourne, Australia

an AUROC of 0.81, indicating a strong ability to distinguish between the two classes and a good agreement between the model's predictions and actual observations. The regression results further enhance our understanding, with the model achieving an MAE of 3.08 on the PHQ-9 scale, which ranges from 0 to 27. This indicates that the model's depression severity predictions are typically within approximately three points of the actual clinical assessments, showing relatively moderate accuracy in quantifying the severity of depressive symptoms.
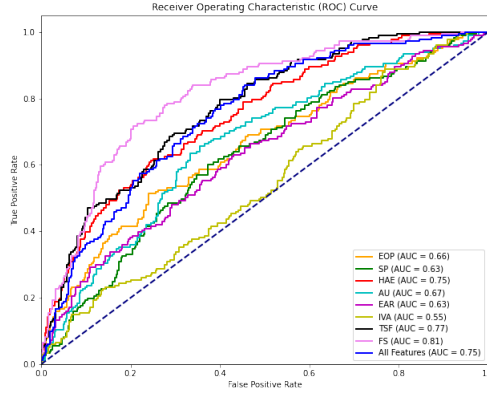

Fig. 5. The ROC plots show the hybrid model performance of each feature type model.

Table 8. Hybrid Model Performance : We trained eight LGBM models for predicting depression, including a different feature subset. The model trained using all features showed the best results in predicting depression

Model	MAE	Accuracy	Precision	Recall	F1	AUROC	No. of Features
Eye Open Probability (EOP)	3.16	0.67	0.50	0.48	0.49	0.66	64
Smiling Probability (SP)	3.26	0.64	0.46	0.43	0.44	0.63	32
Head Euler Angle (HEA)	3.08	**0.72**	0.60	0.51	0.55	0.75	96
Action Units (AU)	3.02	0.67	0.50	0.35	0.41	0.67	384
Eye-aspect ratio (EAR)	3.37	0.64	0.45	0.41	0.43	0.63	64
Inter-vector angle (IVA)	3.57	0.59	0.35	0.28	0.31	0.55	640
Top Significant Features (TSF)	3.18	0.70	0.55	0.55	0.55	0.77	27
Feature Selection (FS)	3.08	0.69	**0.57**	**0.62**	**0.67**	**0.81**	46
All features	**2.81**	0.71	0.59	0.39	0.47	0.75	1280

Overall, the performance of the models varied, but the one using selected features showed the best results in predicting depression. From the AUROC plot (Fig. 5), we can observe even though model with HEA achieved better results in terms of accuracy of 72%, the model itself is stable when combined with other features its yields much better results with more predictive performance stabilization.

6.2.3 Minimum number of days needed to produce reliable detection. The AUROC is a metric used to evaluate the performance of a diagnostic test, with values ranging from 0.5 to 1. A value greater than 0.5 is necessary for the test to be meaningful, and an AUROC of 0.7 or above is generally considered

acceptable. In the context of a depression detection model, the performance was better than random guessing on day 1, with an AUROC of greated than 0.5. On day 1, the model's performance improved to a fair level with an AUROC of 62.4%. Remarkably, starting from day 7, the model achieved an acceptable performance with an AUROC of 71.4%. This progression illustrates (Fig 6) a significant enhancement in the model's ability to accurately detect depression over the weeks.
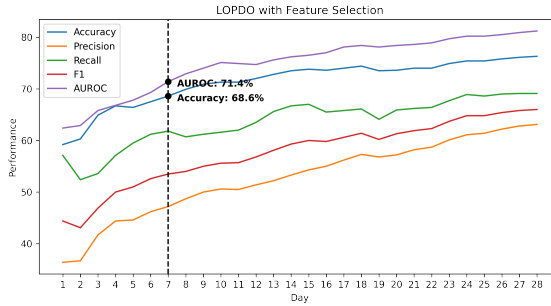

Fig. 6. Minimum number of days needed to produce reliable detection

7   LESSONS LEARNED: DISCUSSING THE DETECTION OF DEPRESSIVE EPISODES ON A SCALE

In this section, we highlight our work that contributes to new insights into the detection of depressive episodes in everyday settings by designing and developing an open-source affective mobile system. We found eye open state, head pose, smile, and action units (2, 6, 7, 12, 15, and 17) as key affective indicators in differentiating depressive and non-depressive episodes that have been validated in our field study. The combined features can be used to predict depression episodes. The universal model has an AUROC of 67%, while the hybrid model has an AUROC of 81%. Further improvement can achieved by collecting more data for subsequent weeks. These results serve as a bridge between controlled laboratory studies and real-world applications, demonstrating the feasibility of depression detection using an affective mobile sensing system.

7.1   Comparison to Prior Work

In the landscape of depression detection using mobile sensing, our approach significantly advances by leveraging in-the-wild data collection through everyday smartphone interactions. This method contrasts with many previous studies that primarily utilize lab-controlled [10, 17, 56, 87, 94] environments for data collection using camera modality, thus limiting the generalizability of their findings. For instance, our work complements the findings of Nepal et al. [66], which also employs an in-the-wild approach but focuses on different feature sets, such as gaze and head pose, along with 2D and 3D facial landmarks. While their study achieves a balanced accuracy of 61%, our method enhances the model's sensitivity to subtler indicators of depression through a detailed analysis of facial action units (AUs), eye-open states, and smile expressions. Our method achieves an AUROC of 81% and Accuracy of 69%, and maintains a consistent MAE across different validation strategies, highlighting the robustness of our findings and their potential for real-world application. This demonstrates the incremental utility of our approach, particularly in the seamless integration of mental health monitoring into daily technology use, thus contributing valuable insights into mobile health (mHealth) technologies for depression detection.

In comparison to mobile sensing, studies focused on depression detection, our approach capitalizes on direct behavioral markers accessible via a smartphone camera, distinguishing it from studies that rely on peripheral sensor data such as GPS or physical activity metrics. For example, Chikersal et al. [15] and Opoku et al. [69] employ a variety of sensors to infer depressive states indirectly through changes in mobility patterns and phone usage. While these studies achieve high accuracy and AUC scores, they may not capture the nuanced emotional states that facial behavior can indicate. Our method's utilization of detailed facial action units, eye-open states, and head gestures offers a more direct and potentially insightful measure of depressive episodes, evidenced by our comparable AUC of 81%. Our model outperformed compared to a model that solely utilized sensor data from the AWARE platform [32], which included Bluetooth (Accuracy=69.3, F1=0.64), Calls (Accuracy=68.5, F1=0.59), GPS (Accuracy=69.5, F1=0.62), and Steps Counter (Accuracy=63.6, F1=0.53), as detailed in research by Chikersal et al. [15]. Our approach achieved an accuracy of 69% and an F1 score of 0.67, surpassing individual sensor results. However, Chikersal et al. reported a higher F1 score of 0.78 when combining all sensors, indicating superior performance compared to our model in that specific setup. This specificity in detecting emotional expressions offers a critical enhancement over traditional mobile sensing methods, making our approach a valuable addition to the spectrum of technologies for monitoring mental health in everyday settings.

7.2 Insights, Challenges and Opportunities in Predicting Depressive Episodes in the Naturalistic Environments

Previous lab-controlled studies have delved into the extraction of facial behavior primitives from face images using affect sensing systems like OpenFace [6]. These studies [10, 17, 56, 87, 94] have achieved impressive performance in detecting depression, due to their high data collection rate (e.g., processing video) and the controlled environment in which the data was collected. In contrast, deploying these systems in real-world scenarios using smartphones results in several challenges. On-device resource limitations often lead to a reduced frame rate (2.5Hz), and the unconstrained data collection settings – affected by factors such as varied lighting conditions, phone orientation, ongoing activities, and whether the environment is indoors or outdoors – can lead to poor quality of data collection. This, in turn, can impact the lack of samples that limits the development of predictive machine learning. Most recently, MoodCapture [66] was introduced that captures facial images in natural environments for depression detection. Their research demonstrated that using a random forest algorithm trained on facial landmarks, it's possible to identify depression and predict PHQ-8 scores among individuals. However, the utility of MoodCapture for developers intending to replicate such studies in different settings may be limited due to the lack of access to the mobile system, dataset, or machine learning framework used by the authors. Our research builds upon MoodCapture's work; we advance data collection by using a literature-based trigger mechanism that responds to user interactions like screen activity and app usage, enhancing user privacy by processing data on-device and discarding raw images post-analysis. Our study is novel in developing an open-source, privacy-aware mobile system that captures and processes facial data in near real-time, introducing significant improvements in privacy-awareness, data collection, and on-device processing.

The similarities and differences observed between lab-controlled and real-world predictive models can be attributed to the inherent nature of the environments in which they operate. Grounded in HCI/affective computing theories [75], controlled environments allow for minimizing external variables [9], ensuring that the subject's emotional state primarily influences the data collected. In such settings, the system can focus solely on the facial behavior primitives without interfering with external factors. However, in real-world scenarios, the myriad of uncontrollable variables, from lighting to personal activities, introduces noise into the data, which can mask or distort the

true emotional indicators. From an affective computing perspective, human-computer interaction is dynamic and multifaceted in real-world settings. The system has to interpret the emotional state and account for the context in which the interaction is taking place. This context can significantly influence the emotional indicators, making them more complex to decipher. These insights are crucial for the design of an affective mobile framework for mental health. Recognizing the challenges of real-world data collection, future frameworks should incorporate adaptive algorithms that can adjust to varying conditions, ensuring consistent and accurate predictions. Additionally, leveraging context-aware computing can help the system contextualize the data, distinguishing between genuine emotional indicators and those influenced by external factors. This approach would lead to a more robust and reliable affective mobile sensing system, enhancing its potential in mental health applications.

Our method demonstrates good performance in depression detection compared to previous work by Opoku et al. [69], which used a similar learning scheme for a hybrid model. However, it's important to note that this comparison may not be entirely direct. Additionally, our universal model shows fair performance compared to prior work by Chikersal et al. [15] in passive sensing using a similar learning scheme. The referenced study incorporates a comprehensive set of features, including sleep patterns, physical activity, phone usage, GPS location, and daily mood ratings, to model behavioral and social signals. However, as extensive research suggests, depression is a multifaceted disorder that affects various aspects of an individual's behavior, social interactions, and physiology. The existing mobile sensing-based approaches [15, 29, 69] have limitations in capturing affective signals [1] manifested through involuntary facial muscle and head gestures, which have been established as crucial indicators of depression. Although wearable sensing-based physiological solutions have been proposed, their high deployment costs including extra cost for purchasing and wearing devices present a significant challenge and lack of affect/emotional signals. Previous research in affective computing conducted within controlled laboratory settings [10, 17, 56, 87, 94] has demonstrated significant promise in capturing emotional signals in individuals with depression. However, these studies are constrained in their applicability to real-world deployments due to computational cost, cost of devices, and user efforts to wear extra devices, particularly when continuous monitoring is essential for delivering timely interventions based on signals captured from users. The resolution of these issues remains unexplored. Therefore, our study aims to bridge these gaps by proposing, collecting, and evaluating the feasibility of deploying data using our novel affective mobile computing system in a real-world, naturalistic setting.

7.3 Privacy and Ethical Considerations to Enhance Feasibility in Real-World Settings

Because a camera temporarily captures a user's face on their smartphone for up to 10 seconds before deletion, users may have concerns due to the human perception of 'using a camera,' even though the system does not record any videos. To balance privacy considerations and data quality for modeling, researchers in the HCI community should consider designing user nudging [5, 30]. These nudging, as suggested by researchers in privacy-preserving systems [20], could allow users to push/pull status about system behavior when running in the background using content that includes appropriate information about data collection status in private contexts where they may feel uncomfortable, even when no images are collected/stored.

While our system automatically removes images after near-real-time feature extraction, the concept of passive sensing [20] that underpins our affective mobile sensing framework does not necessarily require a user interface to reduce users' burden; instead, it runs in the background. This is in contrast to active sensing, typically used for manual tasks. Nevertheless, it is essential to empower users with the ability to enable or disable specific functionalities whenever they perceive potential risks, rather than requiring them to exit the study. While there is a trade-off

between privacy and the quantity of data collected, it is crucial for HCI researchers to collaborate in designing mental health tracking systems that enhance user engagement [68], encourage self-reflection [62], motivation [19], and trust [52], as has been well-established in the field of personal informatics. For example, our individual facial behavior and head pose features can provide users with daily happiness percentage scores using facial expression algorithms. This can serve as one of the motivational strategies that could encourage them to reflect on their mental condition and enable them to maintain good mental health practice, potentially contributing valuable data for long-term healthcare research in the field of mental health.

To balance privacy considerations and data quality during system design, researchers should consider how data could be collected, used, and stored, as well as what implications this could have for the privacy of the users. While there may be discussion over how data processing can dictate/drive the level of invasiveness of the application when the users are given an option of choosing which type of processing or filters (presenting blurry face images for facial feature extraction) [20] they would allow the developers to carry out. At the same time, this necessary allowing the user to do so may impact the number of signals coming from the user's facial data for accurate behavior modeling.

We highlight the high monetary costs associated with wearable-based physiological markers and the lack of rich emotional data. While we agree with the observation regarding the financial aspect, we would want to adequately address the potential privacy costs of camera-based sensing, which collects user data opportunistically throughout the day. The privacy cost in this context refers to the potential risk of unauthorized access or misuse of personal and sensitive visual data captured by the cameras. A method to reduce this impact could involve discarding the user images immediately after computing the sensing values, thereby minimizing the storage of potentially sensitive information or performing the data processing in memory. Further, we would reduce the time for automatic feature extraction (currently 10 seconds per image). While we also provided notification to users that FacePsy is collecting data in the background, more comprehensive and clear justification that considers monetary and privacy costs and strategies to mitigate potential concerns. Facial data of a person contains various characteristics of the face such as shape (face landmarks, smile probability), eye shape (eye-aspect ratio), muscle movements (Action Units), and face orientation (head Euler angles). While these characteristics describe the state of the face at any given moment but don't reveal a person's identity. It is important to ensure that any facial data gathered is encrypted and securely stored, as well as ensure that a user's identity is not revealed in any way. A utilitarian approach from the normative ethics point of view [80] to privacy-maintaining interactions with such data is to balance the benefits that can be gained from facial recognition technology with the potential privacy risks. This approach involves considering the trade-offs between the benefits of using facial recognition technology and the potential risks to privacy. This approach requires making decisions that prioritize the greater good, such as deciding to collect only the minimum amount of data necessary, processing the data in a way that does not reveal user identity, and encrypting and securely storing the data. Additionally, this approach may include taking steps to ensure that facial recognition technology is used responsibly and ethically, such as providing users with information on how their data is being collected and used and allowing them to opt out of the system if they choose to.

7.3.1 User Feedback. We collected feedback (questionnaire available in Appendix A.2) from participants at the end of the study, providing valuable insights into their experiences and perceptions regarding the FacePsy app. Here, we explore the initial reactions, adjustments to the data collection processes, and the evolving acceptance of privacy measures throughout the study. This feedback is instrumental in understanding the real-world implications of deploying such technology and

informs potential enhancements to improve user experience and trust. Below are detailed reflections across five key areas, supported by direct quotes from the participants, illustrating the nuanced reactions and suggestions for future development.

- **Initial Perceptions and Consent**: Participants initially had mixed feelings about facial data collection. For instance, P10 stated, *"I was a little skeptical my apps might hung due to this new functionality."* Despite initial hesitations, the consent process was generally found to be reassuring. P8 shared, *"Yes, everything was properly addressed and I was well informed,"* reflecting a sentiment that the privacy and data usage concerns were adequately managed. Some participants (P23, P30) showed an initial discomfort with the app collecting their sensitive facial data regarding privacy. This discomfort may arise due to the introduction of new technology which they haven't used previously. While discomfort was mentioned by participants, they showed a shift in their comfort level using FacePsy over a period of use.

- **Experience with Data Collection Triggers**: The experience of the app activating on phone unlocking or opening trigger apps varied. While some participants adjusted over time, others remained concerned. P35 commented, *"It definitely became more acceptable over time. Even though I wasn't particularly concerned about data collection, I got more used to it after the first few days. "* showing a quick adaptation, while P24 noted, *"It was a bit of an adjustment but became normal with time"*, showing slow adoption of apps trigger data collection in their daily smartphone use. Whereas P8 noted, *"It was a slight concern at the start but I got used to it eventually."*

- **Impact on Daily Use and Privacy**: Changes in usage habits due to the app's data collection were minimal. The feature that automatically discarded images after 20 seconds was positively received. For instance, P24 appreciated this feature, saying, *"Increased comfort level"*, P23 mentioned *"A bit of comfort, but I am also aware that social media companies convert the images into metadata. where the original image is no longer needed"*. This shows participants were educated about such data collection methods. While P35 mentioned, *"It may have subconsciously led me to use my phone less frequently at beginning. But I got used to over time and got back to my normal phone usage habits. It didn't make me avoid using any specific apps."*

- **Understanding and Trust on-device Feature Extraction**: Understanding of how facial features were extracted varied, with some participants expressing a desire for more information. Trust was generally high for those who felt adequately informed. For example, P30 affirmed, *"I trust it since it was made clear by the researchers that this was the case. I think as far as trust at the data collection, this is a great approach. As long as the image data is not leaving my phone, I do not have any issues with it."*. P35 showed increased comfort level with using their phone over time, noting *"This was the main reason why I became more and more comfortable using my phone after the initial period. I think this was the key feature that made me stick with the study until the end."*. P23 advocated for a mechanism to create a private repository, noting *"Create a private repository and also alert when the data is accessed with providing the reason and by whom it was accessed"* – regarding the ownership and control of their personal data.

- **Long-term Acceptance**: Perceptions of the app improved over time for many participants. Regarding suggested improvements, P8 recommended, *"I would make the activation not so random but in timed intervals"* calling for enhanced user control over the data collection processes. A similar sentiment was echoed by P35, *"Giving users personalization options for which apps to exclude for data collection would be beneficial. Some users may prefer excluding certain apps from being monitored, and I think having this option would increase user acceptance."*

7.4 Potential to Integrate Affective and Cognitive Inferences From Facial Behavior Markers with Conversational Artificial Intelligence (CAI)

As previous research has confirm that understanding facial behavior primitives can enhance long-term solutions for managing depression and provide insights into emotional communication [42], aggression and negative affect recognition [34], psychological distress [89], and automatic thoughts perception during cognitive behavioral therapy (CBT) [82, 83], inferences drawn from facial behavior primitives in our study could enable more effective affective interactions [18, 47] with a virtual CBT agent [82, 83], depending on an individual's emotional and cognitive state. Such an approach aims to create more affect-sensitive multimodal human–computer interactions [73], enhancing the human-like qualities, effectiveness, and efficiency of virtual agents. Real-time interpretation of complex mental states from facial expressions and head gestures, indicating concentration, fatigue, disagreement, interest, contemplation, uncertainty, and more [26], could provide contextual affect data to virtual agents, facilitating more fluid interactions. This aspect is often lacking in text-based and avatar-based therapy agents.

In addition to its potential in depression detection, facial behavior primitives could find application in real-time intoxication detection, as demonstrated in previous work on drunk face detection using an offline dataset [65]. By incorporating such capabilities, the FacePsy system could broaden its utility and use cases for real-world applications. Moreover, the system's applicability could extend to the detection of other forms of intoxication, such as marijuana intoxication [4, 16], where observable facial muscle and head movements are indicative of psychomotor retardation and agitation. Recent developments, like the creation of an augmented reality feedback system for facial paralysis in a mobile setting [7], hint at the potential utility of the FacePsy system in similar deployments. Beyond health-related applications, the FacePsy system's usefulness could be explored in monitoring student engagement in smartphone-based education or virtual classes [45] where instructors can adjust lecture materials based on students' behavioral feedback. Indicators such as concentration, interest, or uncertainty are crucial for enhancing classroom engagement and delivering high-quality education. The FacePsy system represents a significant step in the realm of mobile affect sensing in human-computer interaction, with far-reaching implications for mental health with virtual therapy sessions, substance abuse detection, and educational engagement.

8 LIMITATION AND FUTURE WORK

Even though we were able to get valuable insights about modeling depression and the proposed subset of facial and physiological signals, there are still improvements to be made for this system to be applicable in clinical settings. Although we successfully built a population model to detect depression, however, there might be individual patterns that our population model cannot capture, and thus may limit the generalizability of our model. In our future work, we will collect a larger dataset per participant and investigate the use of more personalized individual models. While we only register 11 categories of app use there could be more categories of app use that could work as the best avenue for data collection, further research should examine if such categories exist. The current limitation of our app, FacePsy, lies in its inadvertent triggering of data collection during intra-app navigation, such as moving between pages within the same app, leading to multiple data captures in a single session. In future work, we plan to refine the app's architecture to discern and limit data collection to significant user interactions, thereby enhancing the efficiency and relevance of the data collection process. Furthermore, in our future work, we want to integrate the pupillary response measurement module [48] in our processing pipeline for in-app measurement of pupillary response by using Android Native libraries. In addition, we aim to enrich FacePsy by integrating it with systems like AWARE and Fitbit, combining rich emotional signals extracted from visual

data with other data such as GPS, heart rate, and EDA. We also plan to add contextual layers, like categorizing apps that trigger data collection in model development which we collect as part of the FacePsy triggering mechanism, to deepen the understanding of facial behavior in context.

Our depression labeling strategy may raise concerns regarding the data's clarity and characteristics: Each session represents a unique data point with specific features. It's important to note that a single day could encompass multiple sessions. However, each session linked to a particular user will predict the different depressive episode labels span over two weeks observation period. This methodology presents two potential challenges. Firstly, the variability between sessions might make it challenging to identify a consistent pattern, which could affect the accuracy of predicting depressive episodes. Secondly, if sessions appear too homogenous, it could suggest that the model might be detecting implicit user characteristics rather than their actual risk of depression.

9 CONCLUSION

Depression is a major mental health disorder. As such, detecting depression can have significant impacts across several domains. Towards this goal, we propose an affective mobile system that allows us to collect facial behavior primitives from faces by opportunistically capturing user faces by observing the user interaction with their phone in a naturalistic setting. To this end, we build a depressive episode prediction model that achieves 81% of AUROC. Our regression model that estimates PHQ-9 scores reached a moderate level of accuracy, exhibiting an MAE of 3.08. Based on the results from our cross-validation, we found our model produces reliable performance from several weeks of data to detect depressive episodes. We highlight key behavior primitives differentiating depressive and non-depressive episodes and use case scenarios regarding how the system could be applicable in detecting mental and neurological disorders for researchers and stakeholders. Lastly, we discuss privacy and ethical considerations in deploying such a system.

10 ACKNOWLEDGMENTS

REFERENCES

[1] Saeed Abdullah and Tanzeem Choudhury. 2018. Sensing technologies for monitoring serious mental illnesses. *IEEE MultiMedia* 25, 1 (2018), 61–75.

[2] Takashi Abe. 2023. PERCLOS-based technologies for detecting drowsiness: current evidence and future directions. *Sleep Advances* 4, 1 (2023), zpad006.

[3] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. [n. d.]. Head Pose and Movement Analysis as an Indicator of Depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (2013-09). 283–288. https://doi.org/10.1109/ACII.2013.53 ISSN: 2156-8111.

[4] Sang Won Bae, Tammy Chung, Rahul Islam, Brian Suffoletto, Jiameng Du, Serim Jang, Yuuki Nishiyama, Raghu Mulukutla, and Anind Dey. 2021. Mobile phone sensor-based detection of subjective cannabis intoxication in young adults: A feasibility study in real-world settings. *Drug and alcohol dependence* 228 (2021), 108972.

[5] Rebecca Balebako and Lorrie Cranor. 2014. Improving app privacy: Nudging app developers to protect user privacy. *IEEE Security & Privacy* 12, 4 (2014), 55–58.

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[7] Giuliana Barrios Dell'Olio and Misha Sra. 2021. FaraPy: An Augmented Reality Feedback System for Facial Paralysis using Action Unit Intensity Estimation. In *The 34th Annual ACM Symposium on User Interface Software and Technology.* 1027–1038.

[8] Alexandre Benoit and Alice Caplier. 2005. Hypovigilence analysis: open or closed eye or mouth? Blinking or yawning frequency?. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* IEEE, 207–212.

[9] Donald T Campbell and Julian C Stanley. 2015. *Experimental and quasi-experimental designs for research*. Ravenio books.

[10] Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. 2023. Depression Recognition using Remote Photoplethysmography from Facial Videos. *IEEE Transactions on Affective Computing* (2023).

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[12] Iwona Chelminski, F Richard Ferraro, Thomas V Petros, and Joseph J Plaud. 1999. An analysis of the "eveningness–morningness" dimension in "depressive" college students. *Journal of affective disorders* 52, 1-3 (1999), 19–29.

[13] Qian Cheng, Wuhong Wang, Xiaobei Jiang, Shanyi Hou, and Yong Qin. 2019. Assessment of driver mental fatigue using facial landmarks. *IEEE Access* 7 (2019), 150423–150434.

[14] Yulia E Chentsova-Dutton, Jeanne L Tsai, and Ian H Gotlib. 2010. Further evidence for the cultural norm hypothesis: positive emotion in depressed and control European American and Asian American women. *Cultural Diversity and Ethnic Minority Psychology* 16, 2 (2010), 284.

[15] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G Creswell, Jennifer Mankoff, J David Creswell, et al. 2021. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 1 (2021), 1–41.

[16] Tammy Chung, Sang Won Bae, Eun-Young Mun, Brian Suffoletto, Yuuki Nishiyama, Serim Jang, and Anind K Dey. 2020. Mobile assessment of acute effects of marijuana on cognitive functioning in young adults: observational study. *JMIR mHealth and uHealth* 8, 3 (2020), e16240.

[17] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–7.

[18] Cristina Conati, Stacy Marsella, and Ana Paiva. 2005. Affective interactions: the computer in the affective loop. In *Proceedings of the 10th international conference on Intelligent user interfaces*. 7–7.

[19] Sunny Consolvo, Katherine Everitt, Ian Smith, and James A Landay. 2006. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 457–466.

[20] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2377–2386.

[21] Abhinav Dhall and Roland Goecke. 2015. A temporally piece-wise fisher vector approach for depression analysis. In *2015 International conference on affective computing and intelligent interaction (ACII)*. IEEE, 255–259.

[22] Sathyanarayanan Doraiswamy, Amit Abraham, Ravinder Mamtani, and Sohaila Cheema. 2020. Use of Telehealth During the COVID-19 Pandemic: Scoping Review. *J Med Internet Res* 22, 12 (1 Dec 2020), e24087.

[23] Naomi Eigbe, Tadas Baltrusaitis, Louis-Philippe Morency, and John Pestian. 2018. Toward visual behavior markers of suicidal ideation. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 530–534.

[24] Paul Ekman. 2003. Emotions revealed: recognizing faces and feelings to improve communication and emotional life. New York. *NY: Times books* (2003).

[25] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[26] 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.

[27] Heiner Ellgring. 2007. *Non-verbal communication in depression*. Cambridge University Press.

[28] 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.

[29] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE wireless health (WH)*. IEEE, 1–8.

[30] Adrienne Porter Felt, Serge Egelman, and David Wagner. 2012. I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*. 33–44.

[31] Jialing Feng, Zhexiao Guo, Jun Wang, and Guo Dan. 2020. Using eye aspect ratio to enhance fast and objective assessment of facial paralysis. *Computational and mathematical methods in medicine* 2020 (2020).

[32] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.

[33] Hans-Ulrich Fisch, Siegfried Frey, and Hans-Peter Hirsbrunner. 1983. Analyzing nonverbal behavior in depression. *Journal of abnormal psychology* 92, 3 (1983), 307.

[34] Siska Fitrianie and Iulia Lefter. 2023. On Head Motion for Recognizing Aggression and Negative Affect during Speaking and Listening. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 455–464.

[35] Wolfgang Gaebel and Wolfgang Wölwer. 2004. Facial expressivity in the course of schizophrenia and depression. *European archives of psychiatry and clinical neuroscience* 254 (2004), 335–342.

[36] Mihai Gavrilescu and Nicolae Vizireanu. 2019. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors* 19, 17 (2019), 3693.

[37] Jean-Guido Gehricke and David Shapiro. 2000. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research* 95, 2 (2000), 157–167.

[38] Simon Gilbody, David Richards, Stephen Brealey, and Catherine Hewitt. 2007. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of general internal medicine* 22 (2007), 1596–1602.

[39] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, Seyedmohammad Mavadati, and Dean P Rosenwald. 2013. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[40] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing* 32, 10 (2014), 641–647.

[41] Google. [n. d.]. ML Kit. https://developers.google.com/ml-kit

[42] Zakia Hammal and Jeffrey F Cohn. 2014. Intra-and interpersonal functions of head motion in emotion communication. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*. 19–22.

[43] Qian-Bei Hong, Chung-Hsien Wu, Ming-Hsiang Su, and Chia-Cheng Chang. 2019. Exploring Macroscopic and Microscopic Fluctuations of Elicited Facial Expressions for Mood Disorder Classification. *IEEE Transactions on Affective Computing* 12, 4 (2019), 989–1001.

[44] Jin Huang and Charles X Ling. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17, 3 (2005), 299–310.

[45] Mohammad Rahul Islam and Sang Won Bae. 2023. *MicroFlow: Advancing Affective States Detection in Learning Through Micro-Expressions*. Technical Report. EasyChair.

[46] Rahul Islam, Karan Ahuja, Sandip Karmakar, and Ferdous Barbhuiya. 2016. SenTion: A framework for sensing facial expressions. *arXiv preprint arXiv:1608.04489* (2016).

[47] Rahul Islam and Sang Won Bae. 2023. Revolutionizing Mental Health Support: An Innovative Affective Mobile Framework for Dynamic, Proactive, and Context-Adaptive Conversational Agents. *Ubicomp, GenAI4PC Symposium* (2023).

[48] Rahul Islam and Sang Won Bae. 2024. PupilSense: Detection of Depressive Episodes Through Pupillary Response in the Wild. *International Conference on Activity and Behavior Computing* (2024).

[49] Asim Jan, Hongying Meng, Yona Falinie Binti A Gaus, and Fan Zhang. 2017. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems* 10, 3 (2017), 668–680.

[50] S Jennifer, Benjamin R Brady, Mohab M Ibrahim, Katherine E Herder, Jessica S Wallace, Alyssa R Padilla, and Todd W Vanderah. 2024. Co-occurrence of chronic pain and anxiety/depression symptoms in US adults: prevalence, functional impacts, and opportunities. *Pain* 165, 3 (2024), 666–673.

[51] Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. 2013. Can body expressions contribute to automatic depression analysis?. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–7.

[52] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. 1–12.

[53] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 593–602.

[54] Ronald C Kessler, Cindy L Foster, William B Saunders, and Paul E Stang. 1995. Social consequences of psychiatric disorders, I: Educational attainment. *American journal of psychiatry* 152, 7 (1995), 1026–1032.

[55] Dimitrios Kollias and Stefanos Zafeiriou. 2021. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792* (2021).

[56] Xinru Kong, Yan Yao, Cuiying Wang, Yuangeng Wang, Jing Teng, and Xianghua Qi. 2022. Automatic Identification of Depression Using Facial Images with Deep Convolutional Neural Network. *Medical Science Monitor: International*

Medical Journal of Experimental and Clinical Research 28 (2022), e936409–1.

[57] Dawn C Kroencke, Sharon G Lynch, and Douglas R Denney. 2000. Fatigue in multiple sclerosis: relationship to depression, disability, and disease pattern. *Multiple Sclerosis Journal* 6, 2 (2000), 131–136.

[58] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine* 16, 9 (2001), 606–613.

[59] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General hospital psychiatry* 32, 4 (2010), 345–359.

[60] Eugene Laksana, Tadas Baltrušaitis, Louis-Philippe Morency, and John P Pestian. 2017. Investigating facial behavior indicators of suicidal ideation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 770–777.

[61] Scott A Laurenzo, Randy Kardon, Johannes Ledolter, Pieter Poolman, Ashley M Schumacher, James B Potash, Jan M Full, Olivia Rice, Anna Ketcham, Cole Starkey, et al. 2016. Pupillary response abnormalities in depressive disorders. *Psychiatry research* 246 (2016), 492–499.

[62] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.

[63] Caio Bezerra Souto Maior, Márcio José das Chagas Moura, João Mateus Marques Santana, and Isis Didier Lins. 2020. Real-time classification for autonomous drowsiness detection using eye aspect ratio. *Expert Systems with Applications* 158 (2020), 113505.

[64] David Matsumoto and Hyi Sung Hwang. 2011. Evidence for training the ability to read microexpressions of emotion. *Motivation and emotion* 35 (2011), 181–191.

[65] Vineet Mehta, Sai Srinadhu Katta, Devendra Pratap Yadav, and Abhinav Dhall. 2019. DIF: Dataset of perceived intoxicated faces for drunk person identification. In *2019 International Conference on Multimodal Interaction*. 367–374.

[66] Subigya Nepal, Arvind Pillai, Weichen Wang, Tess Griffin, Amanda C Collins, Michael Heinz, Damien Lekkas, Shayan Mirjafari, Matthew Nemesure, George Price, Nicholas Jacobson, and Andrew Campbell. 2024. MoodCapture: Depression Detection using In-the-Wild Smartphone Images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 996, 18 pages.

[67] Matthew K Nock, Irving Hwang, Nancy A Sampson, and Ronald C Kessler. 2010. Mental disorders, comorbidity and suicidal behavior: results from the National Comorbidity Survey Replication. *Molecular psychiatry* 15, 8 (2010), 868–876.

[68] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.

[69] Kennedy Opoku Asare, Isaac Moshe, Yannik Terhorst, Julio Vega, Simo Hosio, Harald Baumeister, Laura Pulkki-Råback, and Denzil Ferreira. 2022. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status:: A longitudinal data analysis. (2022).

[70] Kennedy Opoku Asare, Yannik Terhorst, Julio Vega, Ella Peltonen, Eemil Lagerspetz, and Denzil Ferreira. 2021. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. *JMIR mHealth and uHealth* 9, 7 (2021), e26540.

[71] World Health Organization et al. 2003. Investing in mental health. (2003).

[72] M. O'Brien and F. McNicholas. 2020. The use of telepsychiatry during COVID-19 and beyond. *Irish Journal of Psychological Medicine* 37, 4 (2020), 250–255.

[73] Maja Pantic and Leon JM Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* 91, 9 (2003), 1370–1390.

[74] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, et al. 2020. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry* 11 (2020), 584711.

[75] Rosalind W Picard. 1999. Affective computing for hci.. In *HCI (1)*. Citeseer, 829–833.

[76] Mahesh Krishnananda Prabhu and Dinesh Babu Jayagopi. 2017. Real time multimodal emotion recognition system using facial landmarks and hand over face gestures. *Int. J. of Machine Learning and Computing* 7, 2 (2017), 30–34.

[77] Arcady A Putilov. 2017. State-and trait-like variation in morning and evening components of morningness–eveningness in winter depression. *Nordic journal of psychiatry* 71, 8 (2017), 561–569.

[78] Lawrence Ian Reed, Michael A Sayette, and Jeffrey F Cohn. 2007. Impact of depression on response to comedy: a dynamic facial coding analysis. *Journal of abnormal psychology* 116, 4 (2007), 804.

[79] Babette Renneberg, Katrin Heyn, Rita Gebhard, and Silke Bachmann. 2005. Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry* 36, 3 (2005), 183–196.

[80] Pekka Ruotsalainen and Bernd Blobel. 2020. Health information systems in the digital health ecosystem—problems and solutions for ethics, trust and privacy. *International journal of environmental research and public health* 17, 9

(2020), 3006.

[81] Aven Samareh, Yan Jin, Zhangyang Wang, Xiangyu Chang, and Shuai Huang. 2018. Detect depression from communication: how computer vision, signal processing, and sentiment analysis join forces. *IISE Transactions on Healthcare Systems Engineering* 8, 3 (2018), 196–208.

[82] Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. 2020. Analysis of mood changes and facial expressions during cognitive behavior therapy through a virtual agent. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 477–481.

[83] Kazuhiro Shidara, Hiroki Tanaka, Hiroyoshi Adachi, Daisuke Kanayama, Yukako Sakagami, Takashi Kudo, and Satoshi Nakamura. 2022. Automatic thoughts and facial expressions in cognitive restructuring with virtual agents. *Frontiers in Computer Science* 4 (2022), 762424.

[84] Denise M Sloan, Milton E Strauss, Stuart W Quirk, and Martha Sajatovic. 1997. Subjective and expressive emotional responses in depression. *Journal of affective disorders* 46, 2 (1997), 135–141.

[85] Denise M Sloan, Milton E Strauss, and Katherine L Wisner. 2001. Diminished response to pleasant stimuli by depressed women. *Journal of abnormal psychology* 110, 3 (2001), 488.

[86] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. [n. d.]. Spectral Representation of Behaviour Primitives for Depression Analysis. ([n. d.]), 1–1. Conference Name: IEEE Transactions on Affective Computing.

[87] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. 2020. Spectral Representation of Behaviour Primitives for Depression Analysis. *IEEE Transactions on Affective Computing* (2020).

[88] Siyang Song, Enrique Sánchez-Lozano, Mani Kumar Tellamekala, Linlin Shen, Alan Johnston, and Michel Valstar. 2019. Dynamic facial models for video-based dimensional affect estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[89] Giota Stratou and Louis-Philippe Morency. 2017. MultiSense—Context-aware nonverbal behavior analysis framework: A psychological distress use case. *IEEE Transactions on Affective Computing* 8, 2 (2017), 190–203.

[90] Anja Stuhrmann, Thomas Suslow, and Udo Dannlowski. 2011. Facial emotion processing in major depression: a systematic review of neuroimaging findings. *Biology of mood & anxiety disorders* 1, 1 (2011), 1–17.

[91] Vincent W.-S. Tseng, Saeed Abdullah, Jean Costa, and Tanzeem Choudhury. [n. d.]. AlertnessScanner: what do your pupils tell about your alertness. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Barcelona, Spain, 2018-09-03) *(MobileHCI '18)*. Association for Computing Machinery, 1–11.

[92] Vincent W-S Tseng, Saeed Abdullah, Jean Costa, and Tanzeem Choudhury. 2018. AlertnessScanner: what do your pupils tell about your alertness. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[93] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 3–10.

[94] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 3–10.

[95] Jessie M VanSwearingen, Jeffrey F Cohn, and Anu Bajaj-Luthra. 1999. Specific impairment of smiling increases the severity of depressive symptoms in patients with facial neuromuscular disorders. *Aesthetic plastic surgery* 23 (1999), 416–423.

[96] Sarah Collier Villaume, Shanting Chen, and Emma K Adam. 2023. Age disparities in prevalence of anxiety and depression among US adults during the COVID-19 pandemic. *JAMA network open* 6, 11 (2023), e2345073–e2345073.

[97] Philip S Wang, Gregory E Simon, Jerry Avorn, Francisca Azocar, Evette J Ludman, Joyce McCulloch, Maria Z Petukhova, and Ronald C Kessler. 2007. Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes: a randomized controlled trial. *Jama* 298, 12 (2007), 1401–1411.

[98] Rui Wang, Andrew T Campbell, and Xia Zhou. 2015. Using opportunistic face logging from smartphone to infer mental health: challenges and future directions. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 683–692.

[99] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[100] Fang Zhang, Jingjing Su, Lei Geng, and Zhitao Xiao. 2017. Driver fatigue detection based on eye state recognition. In *2017 International Conference on Machine Vision and Information Technology (CMVIT)*. IEEE, 105–110.

[101] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2018. Visually interpretable representation learning
for depression recognition from facial images. *IEEE Transactions on Affective Computing* 11, 3 (2018), 542–552.

A SURVEY

A.1 PHQ-9

Participants were asked: "Over the last two weeks, how often have you been bothered by the
following problems?" This questionnaire is part of the standard assessment to gauge the severity of
depressive symptoms. Below is the PHQ-9 questionnaire (Table 9) used in the study:

Table 9. PHQ-9 Questionnaire Items

No.	Question
1	Little interest or pleasure in doing things
2	Feeling down, depressed, or hopeless
3	Trouble falling or staying asleep, or sleeping too much
4	Feeling tired or having little energy
5	Poor appetite or overeating
6	Feeling bad about yourself - or that you are a failure or have let yourself or your family down
7	Trouble concentrating on things, such as reading the newspaper or watching television
8	Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual
9	Thoughts that you would be better off dead, or of hurting yourself

A.2 Study Feedback

Table 10 lists the user feedback questions administered at the end of the study to gauge participants'
perceptions of the FacePsy app, focusing on aspects of consent, data collection triggers, impact
on privacy, understanding of feature extraction, and long-term acceptance. The questions were
designed to understand the participants' experiences throughout their interaction with the app and
to gather suggestions for future improvements.

Table 10. User Feedback Questions

No.	Question
1	When you first started using the FacePsy app, what were your initial thoughts about the facial data collection, especially when unlocking your phone or using specific apps?
2	How were you informed about the data collection process, and did you feel that the consent process adequately addressed your concerns about privacy and data usage?
3	Can you describe how you felt the first few times the app activated upon unlocking your phone or opening trigger apps? Did it become more acceptable over time, or did it remain a concern?
4	Were there any particular trigger apps that made you more uncomfortable when the FacePsy app activated? How did this affect your usage of those apps?
5	Did you notice any changes in your phone usage habits due to the app's data collection methods? For example, did you use your phone less frequently or avoid certain apps?
6	What are your thoughts on the app automatically discarding images after 20 seconds? Did this feature influence your comfort level with the ongoing data collection?
7	How well do you understand the process of facial feature extraction by the app? Was there enough information provided about what data is extracted and how it is used?
8	Do you trust that the facial data collected remains on your device and is not uploaded elsewhere? What could increase your trust in the system's handling of your data?
9	How has your perception of the FacePsy app and its data collection practices changed during the course of the study?
10	What improvements or changes would you suggest for the app, especially regarding user control over data collection and privacy?