A Review on Mood Assessment Using Smartphones

Zhanna Sarsenbayeva[1(✉)], Charlie Fleming[1], Benjamin Tag[2], Anusha Withana[1], Niels van Berkel[3], and Alistair McEwan[4]

[1] School of Computer Science, University of Sydney, Sydney, Australia
{zhanna.sarsenbayeva,charlie.fleming,anusha.withana}@sydney.edu.au
[2] Department of Human Centred Computing, Monash University, Melbourne, Australia
benjamin.tag@monash.edu
[3] Department of Computer Science, Aalborg University, Aalborg, Denmark
nielsvanberkel@cs.aau.dk
[4] School of Biomedical Engineering, University of Sydney, Sydney, Australia
alistair.mcewan@sydney.edu.au

Abstract. Due to their abundance of sensors, today's smartphones can act as a scientific tool to collect contextual information on users' emotional, social, and physical behaviour. With the continuously growing amount of data that can be unobtrusively extracted from smartphones, mood-tracking and inference methods have become more feasible. However, this does raise critical implications for end-users, including accessibility and privacy. Following a structured selection process, we reviewed 32 papers from the ACM Digital Library on mood inference and tracking using smartphones. We conducted an in-depth analysis of used sensors, platform and accessibility, study designs, privacy, self-reporting methods, and accuracy. Based on our analysis, we provide a detailed discussion of the opportunities for research and practice that arise from our findings and outline recommendations for future research within the area of smartphone-based mood tracking and inference.

Keywords: Mood tracking · Mood inference · Smartphones

1 Introduction

Our mood has a profound impact on our physical and mental health, as well as our financial, educational, and social well-being [38]. Often, mood and emotion are falsely seen as synonymous. But while there exist many similarities and correlations between mood and emotion [77], they are not identical. Beedie et al. [12] reviewed 65 papers to identify differences between the two. The authors found that 62% of the articles identified duration, 41% intentionality, 31% cause and consequences, and 18% identified function as key distinctive factors. This suggests that while emotion can contribute to mood, mood is more intentional and

J. Abdelnour Nocera et al. (Eds.): INTERACT 2023, LNCS 14143, pp. 385–413, 2023.
https://doi.org/10.1007/978-3-031-42283-6_22

present for a longer duration. Consequently, mood issues can lead to more long-lasting consequences, which is clearly demonstrated via the correlation between mood instability, low mood, and psychiatric issues [71].

Mood tracking and inference applications are more ubiquitous than ever before. As of 2022, 83.32% of the world population owns a sensor-rich smartphone [9]. In 2021, two out of three teenagers and young adults reported that they had previously used a mental health application [76]. According to Caldeira et al. [22], users of publicly available mood tracking and inference apps use them to (1) learn about their idiosyncratic mood patterns, (2) improve their mood, and (3) monitor and manage mental illness. These apps are also often used to track and manage stress [22]. Negative mental health outcomes can be mitigated by such applications, as they provide a tool for increasing self-awareness and enabling early interventions [23].

In this paper, we report on a systematic literature review of the existing ACM literature focusing on tracking and inferring user mood using smartphones. Based on this, we provide an in-depth analysis of used sensors, technology, study methods, privacy details, self-reporting methods, and accuracy. Additionally, we present a quantitative analysis of commonly used sensors for passive mood tracking. We put additional focus on the accessibility of each method according to its availability on different smartphone OSs and external device requirements. We then review study details, quantifying and comparing age, gender split, participation rates, participation rewards, duration, socioeconomic background, and more. Given the limited number of literature reviews within the domain of mood tracking using smart devices [60] and the rapid developments in this field, our work provides an up-to-date overview of how mood tracking is performed and addressed in HCI user studies, particularly using smartphones. Prior literature points to the variety of mood detection systems, including self-reported data, speech, facial recognition, mobile phone usage patterns, and physiological signals, with the authors focusing primarily on the accuracy and usability within each of them [60]. We extend this prior work by adding a novel analysis of the critical and increasingly important privacy and accessibility considerations, as well as sensor prevalence. We then discuss our results with reference to the relevant literature, providing implications for future studies in the area concerning self-report methods, sensor choices, platforms, accessibility, and privacy considerations. To summarise, the contribution of our work is as follows.

1. We provide an overview of related work in the field of *Mood tracking* and *Inference*, including what its current state looks like and how it arrived there.
2. We then outline our method for the literature review, including database choice, search query, and table creation. We provide a summarised version of the analysed data in a table summarising the key metrics.
3. Key metrics across each study are then quantified in a uniform fashion before being compared and contrasted within the results section.
4. Based on inferences gained from our findings, we provide recommendations for future studies in this domain.

2 Related Work

In the following, we provide insights into the state-of-the-art research on the importance of mood tracking using smartphones and methods and sensors commonly used in assessing mood.

2.1 Importance of Mood Assessment

Mood reflects a key component of our health. Hence, tracking, understanding, and interpreting our moods gives us a greater sense of power and control over our lives. Considering clinical application, the capability to track mood helps people identify and mitigate psychiatric disorders [75]. For instance, mood instability is a key diagnostic criterion for bipolar disorder [4]. It is also heavily related to other mental health disorders and is a precursor behind undesirable clinical outcomes, such as borderline personality disorder, bipolar disorder, and depression [71]. Similarly, affective states which are consistently negative have been reported to be linked to the propensity of developing depressive disorders [21]. The impacts of unmediated psychiatric health states are far-reaching, including high school drop-out [49], divorce rates [50], and early parenthood [48].

Additionally, mood significantly influences our cognition, sociability, and productivity. For instance, Mitchell and Phillips [66] showed that small mood fluctuations could significantly impact neural activation and cognition, with implications for our memory and thinking. Bower [19] showed how different mood states can trigger associated memories and experiences and influence our perception and judgement. In a review of consumer behaviour, Gardner [35] mentioned how people in good moods are more likely to care about their future and attempt new things. In their research, Lyubomirsky et al. [63] found that long-term positive affective states ('happiness') correlate with numerous behaviours often understood as synonymous with success.

Toegel et al. [90] mentioned how managers that exhibit strong self-monitoring and positive affect disposition are more likely to provide emotional support to others in the workplace. Carlson et al. [26] demonstrated how positive mood can increase 'helpfulness', among other factors. In the same paradigm, George [36] demonstrated how a positive mood at work is associated with higher sociability & greater sales performance.

The multidimensional impact by which mood influences our daily lives is also demonstrated by the versatility of mood inference and tracking tools. For patients with illnesses such as bipolar disorder or depression, maintaining a stable mood is a goal that can be aided by such tools [69,86,96]. Calear and Christensen [23] found that the negative effects of depression can be mitigated when the symptoms are identified early. Wang et al. [95] implemented a system to track depression in workers, finding that interventions could lower their depression scores and significantly improve job retention. For those with stress disorders, the focus is mainly on identifying stress, its triggers, and mitigating its intensity [17,28]. For a broader public, mood-tracking applications bring benefits

such as enhancing general mood awareness [33], highlighting the importance and foreseeable benefits of mood monitoring.

2.2 Mood Inference

Many different approaches exist within the inference domain. The most common approach towards inferring mood is quantifying affective states proposed in Russel's circumplex model Russell [79]. As do moods, affective states can have drastic impacts on behaviour, cognition, perception, and reflexes [78]. At the core of any mood event is a 'core affect', representing a feeling of 'good' or 'bad', 'energised', or 'enervated'. Users can report their mood using a 2D grid expressing valence (positive or negative affect) on the x-axis and arousal (intensity) on the y-axis. Other common measures of mood include self-report tools, such as the 'Patient Health Questionnaire' (PHQ) score [53], 'Patient Activation Measure' (PAM) [44] and 'Positive and Negative Affect Schedule' (PANAS) [97].

While some methods rely purely on self-reporting as a 'tracker', most methods rely on self-reporting to establish ground truths. Recently, passive sensor data and machine learning (ML) models have been employed to correlate various activities with self-reported mood states [102]. From here, current mood can be inferred, and future moods can be predicted, with predictive strength being dependent on the sensor data used and the accuracy of the model.

There is a low amount of review literature within the domain of mood tracking and inference using smartphones. One example is a survey by Lietz et al. [60], investigating different types of mood detection systems using different modalities, such as self-reports, speech, facial recognition, mobile phone usage patterns, and physiological signals. Lietz et al. [60] provide several robust analyses regarding the efficacy of different mood inference input modes. However, their overview does not consider specialised sub-categories (e.g., systems that use online social networks as a data source). Further, complementary to the review by Lietz et al. [60], we consider study details, technology types, and a series of other metrics to review mood-inference literature.

2.3 Mood Tracking on Smart Devices

Tracking and inferring mood is not a new concept; however, it was previously limited by the requirement for (1) pen-and-paper reporting and (2) psychologists' analysis of the data [102]. As smartphones have become our constant companions and are equipped with many sensors, they can be an almost completely unobtrusive tool to assess and monitor users' moods. Recent global events, such the self-isolation caused by the COVID-19 pandemic, demonstrated a need for unobtrusive mood tracking that enables constant monitoring of users' emotional states to support their mental well-being.

Recently, a myriad of studies have digitalised traditional pen-and-paper reporting methods and (partly) introduced them into the mobile domain [87]. Khue et al. [51] highlighted the benefits of mobile technology as enabling mood

assessment in more naturalistic settings outside the lab, where data can be collected more longitudinally and *in situ* across a variety of scenarios. According to Khue et al. [51], 83% of their 48 participants preferred mobile scales to pen-and-paper scales. The wide adoption of smart devices, integration of clinical methods, and the development of predictive ML models have accelerated the efficacy of mobile methods. For instance, Chan et al. [27] showed how mobile self-reporting is statistically significant with other clinical tools and thus is a valid means of assessing mania and depression symptoms in bipolar patients. More experimental methods use smartphones to passively read affect values off of users' faces [88]. While facial expressions have been disputed as sole sources for affect inference [11], multi-sensory approaches to capture emotion, affect, and mood inference provide promising avenues [103].

However, some challenges do exist in the ubiquitous computing domain. Mehrotra et al. [65] explained how receiving high-quality ESM data is challenging due to respondents not answering honestly or ESM prompts being ignored as they are too frequent and cumbersome. Further, Lee et al. [57] stated that 66% of users that downloaded mood-tracking apps only reported their mood once. Users are also more likely to report mood when it is positive and ignore reporting when in negative mood states [82]. Looking at users that report depression, Depp et al. [30] found that compliance rates were significantly lower for mobile phone reporting compared to pen-and-paper reporting. On the other hand, mobile reporting was able to capture variability and concurrent validity to a greater standard when quantifying affect indicators. Van der Watt et al. [98] showed how depressed patients are willing to use mobile phones to track and assess their mood states. However, without distant support, it may be difficult to mitigate adverse mental health outcomes. On the balance of the evidence, challenges do exist adapting this traditionally formal process into a casual mobile environment. Despite this, the potential benefits make this a rational domain for future research and development [89].

Usage of Sensors for Mood Tracking. In the context of sensor-based mood tracking, it is important to explore the sensors utilised in previous studies and highlight the existing techniques and approaches to provide a comprehensive overview of the existing literature and methodologies employed in the field.

The microphone and accelerometer sensors are ubiquitous and allow for the accurate construction of predictive features based on pertinent real-world information. For instance, Spathis et al. [85] hypothesised that mood is heavily influenced by our activity, environment, and surrounding noise levels. By using the microphone and accelerometer, Spathis et al. [85] were able to collect insights into these three factors, classifying users as 'relaxed' or not with 74% accuracy. Further, Servia-RodrÍguez et al. [83] collected the microphone and accelerometer data of 1,556 and 1,656 users respectively, as well as the call/text log of 8,247 users. Their empirical analysis found that the accelerometer and microphone presented a higher significant correlation with user mood than text/phone logs.

While microphones and accelerometers are considered useful sensor choices for mood prediction, the manner in which they are used and the information derived from them varies between studies. With respect to microphones, a significant reason for their varied usage is due to the privacy concerns that arise from recording and storing microphone data [54]. While Spathis et al. [85] and Bachmann et al. [6] used snippets of amplitude data from the recordings, Zhang et al. [105] extracted the mean, variance, and noise-to-silence ratio of snippets. A common element of all these studies is their use of microphones to infer contextual information about its user.

Conversely, Wang et al. [96] used the microphone in a less general manner by letting it operate continuously. While this raised privacy concerns, data was obfuscated, and the microphone proved highly effective for detecting unique/non-unique conversations and thus quantifying levels of social interaction. Chang et al. [28] also used the microphone but focused on the extraction of speech-related features (e.g., mean, SD, pitch and others). The study demonstrates the predictive power of the microphone sensor with 75% accuracy on two-class emotion classification and 84% accuracy in identifying stressful situations. A similar study by Lu et al. [62] use acoustic features from the microphone to detect stress with 81% accuracy indoors and 76% accuracy outdoors. These examples highlight the wide variety of data and insights that can be obtained using the microphone sensor.

While the general use of microphones has been to infer environment, social, and speech data, all but one study using the accelerometer aimed to track physical activity levels. The exception was [25], where the accelerometer was used to track how users were holding their phones when typing. Most studies use the accelerometer to infer physical activity levels, making it a widely adopted data probe for this purpose [101]. However, data can be challenging to model due to errors i.e., the cross-axis effect and non-linearity of data [3]. Nevertheless, if the accelerometer can be leveraged effectively to capture physical activity levels, mood prediction becomes more accurate overall. This is because a significant correlation exists between physical activity, health, and general well-being [72].

3 Method

The following section outlines the methodology followed to identify and summarise a final set of 36 studies used for our literature review. This includes the choice of the database, construction of the search query, and filtering of results.

3.1 Database Choice

We conducted a literature review on studies that propose mood-tracking and inference systems using smart devices. Our focus pertained to digital mood tracking and inference systems in Human-Computer Interaction (HCI). To obtain a broad overview, we conducted our search in the Association for Computing Machinery's Digital Library (ACM DL). While this means that work published

in non-ACM venues is not included in our review, we argue that the ACM DL provides a comprehensive representation of HCI venues.

3.2 Search Query

To generate a reasonably sized set of works, the following search query was used: *'[[Title: mobile phone mood] OR [Title: mobile phone mood inference] OR [Title: mobile phone mood tracking] OR [Title: mobile mood tracking]] AND [Fulltext: mood]'*. Due to the significant developments in mobile-device technology, we only considered articles from 2010 onwards. The search query returned 216 results, all of which were manually analysed. Papers were excluded from our analysis if they did not offer a mood tracking or inference system or if their system was not applicable to mobile-device usage. Based on these criteria, we retained 32 papers out of 204 for further analysis. The median year of these papers' publication was 2017, and the set included a total of 36 studies. While we note that our scope could be limited to the results of this search and it is possible some papers were not included in our sample, our chosen keywords provided an extensive tailored search space.

3.3 Metric Analysis

A set of 32 papers was transcribed into a table matrix, with each row representing a paper and each column representing a metric. Some papers had multiple studies, resulting in two rows per paper. Initially, metrics for *accuracy*, *usability*, *privacy*, and *study design* were proposed. After analyzing 10–15 papers with these metrics, they were refined and split into more quantitative metrics for further analysis. The final set of metrics included sensor types, number of sensors used, passive tracking, feature extraction, self-reporting methods, accessibility details, usability details, privacy details, and study details. We carefully analysed each paper, extracting relevant information into a spreadsheet. In total, we recorded 38 metrics per paper, encompassing both qualitative and quantitative data. Tables 2, 3, 4, 5 and 6 provide a summary of these metrics, excluding qualitative metrics due to space limitations, although they are used in subsequent analyses.

4 Results

A total of 36 studies originating from 32 papers were analysed with respect to sensor details, study details, technology details, privacy details, and Experience Sampling Method (ESM) strategy. For the sake of analysis, only papers that included a formal study were included below. The resulting data is outlined below with subsequent discussion included in the following section.

4.1 Sensors and Phone Use

We began by compiling the number of smart device sensors employed within each paper. From here, it was found that 22 out of 36 studies (61%) used at least one type of sensor to derive passive mood predictions. A total of 16 studies (44%) used two or more, 12 (33%) used three or more, and nine (25%) used four or more. The papers that used the greatest number of sensors were [69,96] with 12, both making use of the StudentLife system [96]. Conversely, there were 10 papers that made use of just one sensor. Of the studies that used at least one sensor ($N = 22$), the average number of sensors used was four and the median was three. The two most commonly used sensors are the accelerometer (12) and microphone (12). This is closely followed by application usage (10), SMS/Call info (9), Wi-Fi sensor (9), and GPS (8). Additionally, the light sensor was used in six studies, a screen on-off in five studies, Bluetooth in four studies, a calendar in two studies, and a keyboard in two studies. Of the 12 studies used **activity tracking**, four studies (33%) tracked activity via a physiological sensor (e.g., smartwatch). The remaining eight studies (67%) tracked activity using phone-only sensors i.e., accelerometer or GPS.

Feature selection is a complex process that involves carefully analysing and selecting the data that gets input into predictive models. An explicit feature selection process has important implications for future studies that incorporate predictive machine learning models. Hence, the proportion of studies that highlighted a feature selection process was outlined. Of the 36 studies, 27 (76.3%) were explicit about having a feature selection process. The nine studies that did not include a feature selection process did not contain any passive tracking components.

4.2 Study Design

Our analysis of the study design included the average number of participants, socioeconomic background of participants, type of study, gender balance, age of participants, duration of the study, participant rewards, and the number of studies per paper. Across 36 studies that reported participant information, the average number of participants was 1,483. However, this figure is skewed by a few studies that used large data sets compiled from general public app usage. We, therefore also report the median number of participants being 34.

Our analysis determined that 42% of the studies conducted had 0–25 participants. 17% of studies employed 26–50 participants, and another 17% employed 51–100 participants. It is worth noting that 75% of all studies employed between 0 and 100 participants. Conversely, 11% of studies employed between 101–500 participants, and 14% of studies employed over 1000 participants. Of the five studies that included over 1,000 participants, four studies leveraged back-end access to public applications to compile participant data. The other study operated on public Twitter data via API scraping. The study with the largest participant cohort was by Servia-Rodríguez et al. [83] with 18,000 participants. The study with the least participants was by Lietz et al. [59] with two participants.

Each study was categorised as an 'in-the-wild' or a 'lab' study. Out of 36 studies, 29 (81%) were conducted in-the-wild, 6 (17%) were conducted in a laboratory setting and one study did not specify. Out of 32 papers, four (10%) papers contained multiple studies. Furthermore, 31 of 36 studies (86%) were explicit about the duration of their study. For the purpose of analysis, the study duration was split into bins of 1 day, 3–4 days, 11–14 days, 4 weeks, 5–6 weeks, 8 weeks, 18–28 weeks, 36 weeks, 52 weeks, and greater than 52 weeks.

Demography. Out of 36 studies, 21 (57.9%) explicitly reported on the **socioeconomic background** of their participants. An additional seven studies were identified as 'Random Mixed' due to operating on usage data from a public application or online social network (OSN). This meant that the socioeconomic background of participants could be inferred within 28 (78%) studies. From here, we identified the societal group which was targeted most amongst the study set – university students. A summary of this information is provided in Table 1.

The next study parameter that was analysed was the **distribution of gender**. Of the 36 applicable studies, 18 (50%) studies reported the distribution of gender by identifying the number of male and female participants. An additional six (33%) studies were identified as 'Random Mixed' due to operating on large data sets compiled from public app usage or OSN activity. One study operated on a large data set, however, specified that their application was targeted mostly towards women (no specific proportion given) [18]. For the sake of analysis, only the studies that reported their gender split were considered.

The next stage of analysis involved assessing the average number of female and male participants across the set of applicable studies. Among the 18 studies that reported gender as a male/female split, study participants were 49.2% female on average, with males making up 50.8%. Four of the 18 studies (22%) had under 35% female participation. At an aggregate level, half of the studies that reported gender had under 46% female representation. Conversely, three of the 18 studies (17%) had under 35% male participation. One additional study had under 46% male representation. As an aggregate, 22% of the studies that reported gender had under 46% male representation.

4.3 Mobile Technology

A total of 36 studies were analysed with respect to the type of phone OS that was used as well as the type of smart device (s) required. With implications for **accessibility**, we then analysed the proportion of mood inference applications that were 'Android only', 'iPhone only', or available on all smartphones. From 36 studies, 30 (83%) provided details about which phone type the proposed application could be run on. From here, results were placed into categories based on the most commonly used phone types.

From our analysis, we can observe a significant split in the distribution of application offerings per phone type, with a clear preference for Android. 20 of these 30 studies (66%) that reported OS details used an application that was only accessible via Android OS. Conversely, two studies (7%) used an application that was only available via iOS. From here, a clear history for an application

offering to cater to one phone type is identified. Only six studies (20%) offer options on both Android/iOS and only two studies (5.9%) use software that can be used on all smartphones Chang et al. [28].

Acknowledging additional implications for accessibility, the next stage of analysis pertained to whether the study required: (1) Phone only, (2) Phone + Wearable, (3) Phone + OSN, (4) Wearable only or (5) OSN only. All 36 applicable studies reported details about the required technologies.

Our analysis demonstrates that 25 of 36 studies (69%) required only a smartphone as a single data source. This adds further evidence to the convenience and accessibility of harnessing the phone and its sensors for mood inference goals. Conversely, two studies used a wearable smart device as their primary source of data, with an additional two studies leveraging online social network data as their single source. While 29 studies (81%) used a single source of data, the remainder used a combination of the aforementioned data sources. Four studies (11%) required a wearable sensor in addition to a smartphone. A further four studies (11%) required a smartphone and active online social network presence. It must also be noted that one study [8] required an external LCD screen in addition to a phone.

4.4 Privacy

With most studies reporting on the usage, storage, and analysis of sensitive personal data, privacy is an important consideration. We began by assessing the studies' privacy-related practices according to the below criteria.

With each study operating on personal data, we first considered whether they employed any privacy-preserving methodology. A majority of 27 (75%) out of 36 studies address privacy concerns by either hashing or encrypting personal data, recording microphone amplitudes rather than raw data, using non-GPS location measures, categorising idiosyncratic data (e.g., app usage), using public data, or hosting data locally or in highly secure environments. Conversely, 9 (25%) studies do not mention that they perform any of the aforementioned privacy-preserving procedures.

4.5 Self-reporting

The papers were analysed with regard to their usage of Experience Sampling Methods (ESM)/Ecological Momentary Assessment (EMA, used interchangeably). Our analysis extended towards the proportion of studies that used ESM approaches, ESM, and passive tracking approaches, as well as passive tracking only. From 36 applicable studies, 9 (25%) used some form of experience sampling to derive self-reported mood data from their participants while not employing any passive mood-tracking capabilities. We found that 21 (58%) studies used ESM and passive tracking with at least one sensor. Finally, six (17%) of the studies purely relied on passive tracking to infer mood. A total of 24 studies explicitly reported how often they prompted users to report their mood. Table 1

outlines the frequency of daily mood input and the number of corresponding studies.

From the 30 studies that employed mood scales and logging, 15 (50%) utilised the Affect Grid based on Russel's Circumplex Model of Affect [79]. Other strategies include PAM [44] as used in four (13.33%) studies, a valence-only mood scales as used in four (13.33%) studies, PANAS [97] as used in three (10%) studies, Ekman's discrete category mood model as used in two (6.67%) studies, SAM [20] as used in one (3.33%) study, SPANE [32] as used in one (3.33%) study, AffectButton as used in one (3.33%) study, Mood category categorisation as used in one (3.33%) study, EPDS as used in one (3.33%) study, and a 2-item assessment of positive/negative activation created as used in one (3.33%) study. It is important to note that AffectButton, PANAS, PAM, and SAM methods are derived from the Circumplex Model of Affect. Additionally, scales i.e., PHQ score [53] and GAD-7 were used in four (13.33%) studies respectively.

5 Discussion

5.1 Mood Inference

Self Reports. Perhaps the most valuable piece of information amongst mood inference strategies is the output derived from user mood reporting. This output can be placed into a timeline such that it can be viewed empirically over time, or directed into an ML model to predict future moods. As mood tracking has moved into the digital domain, mood assessment has moved from relying on self-reports collected in clinical settings to focusing on experience sampling. This has been an effective strategy as ESMs are able to limit recall bias, which is highly prevalent in clinical settings [84]. Further, ESMs also facilitate the analysis of micro-events that drive real-world behaviours [13]. This means that context-dependent events that trigger mood changes can be quantified.

While collecting context-dependent participant data in real-time has significant advantages, participant non-compliance can quickly downplay these advantages. With multiple inputs usually required per day, and some inputs taking excessive amounts of time, this process is often burdensome to users. As a result, the compliance rates are generally substandard [99].

Across 24 studies that used mobile ESMs, compliance reduced from 91.7% to 77.4% as prompting frequency grew from 2–3 times to 4–5 times per day [14,99]. On average, similar rates of compliance exist in our study set. Of 24 studies that required daily mood reporting, 18 (50%) studies required 1–3 mood reports per day, four (11%) studies required 4 per day and two studies (6%) required 12 per day. Of the studies that required 1–3 mood inputs per day, compliance rates were higher than expected.

However, key variances were identified according to idiosyncratic study processes. Khue et al. [51] was able to achieve a 98% compliance rate prompting 2x per day. Li and Sano [58] had a 93.7% compliance rate prompting 1x per day with 3 short 0–100 scale inputs for mood, health, and stress. Lee et al. [56] monitored 36 participants and received a compliance rate of 88% prompting once per

day using Affect Grid and PANAS. While compliance rates are higher on average in this lower prompting frequency range, there are some exceptions. Torkamaan and Ziegler [91] recruited 547 users and prompted twice a day, however, only 391 (71.5%) users met the required threshold of at least three mood entries over a seven-day period. Visuri et al. [93] used two data sets containing a total of 36 people and prompted twice per day, however, only 61% of participants completed enough self-reports (over 20) to be included.

The general trend continues when analysing studies that prompt users at higher frequencies. Zhang et al. [106] prompted users three times per day using a discrete 1–5 rating of 6 basic Ekman emotions and recorded a participation rate of 77.3%. Similarly, Zhang et al. [105] prompted three times per day using the same method and found that 71.4% of participants had enough entries for their data to be included. Wang et al. [96] prompted four times per day using a PAM scale, recording a participation rate of 80%. LiKamWa et al. [61] required four mood inputs per day using the affect grid, however only 75% of participants managed to provide enough entries.

While a negative correlation between prompting frequency and compliance rates is demonstrated in our study set, this clearly is not the only contributing factor. There is evidence that dropout rates are not purely based on frequency, but also on the complexity of mood input. Alvarez-Lozano et al. [2] required just one mood input per day, however, the mood report included four scales, five yes/no questions, and three numerical inputs. While this study ran for five months, they were not able to access large amounts of data for most users due to instances where self-assessment tests weren't provided. This corroborates the idea that regular mood reporting must be an efficient process, or else users will feel burdened.

Some studies [91,93] also recorded poor participation rates, and only prompted users twice per day. Hence, even if reporting frequency is lower, we observe that an overly complex mood reporting process will lead to non-compliance. Bond et al. [18] mentions low compliance with mood scales GAD-7, PHQ-9, and EPDS due to these more complex scales being slow and mentally arduous. In contrast, mood logs are more efficient to complete and require less cognitive workload [18]. With a ratio of 3.28 mood logs per mood scale completion in this study, it indicates that participants prefer to report their mood through an efficient affect-grid-based ESM rather than a complex mood scale. Bond et al. [18] concluded that users prefer simple efficient ESMs, demonstrating that a shift in user behaviour occurs when complex ESMs are used instead of simple ones. Wallbaum et al. [94] underwent a field study with 18 participants, receiving preference data for reporting methods PAM, SAM, emotion terms (PAM but word-based), and colour input (colour indicates mood).

However, qualitative data revealed that users didn't prefer any method. Instead, they liked the option of having efficient and alternative ways to input their mood according to situation and context. Hence, future mood studies should consider adding multiple input method options to encourage reporting across a variety of scenarios. This should include efficient measures in tandem

with more verbose measures. While this may increase complexity, it will likely reduce data inconsistency. The evidence suggests that, in order to maintain high rates of compliance, a careful balance between reporting strategy, flexibility, efficiency of input, and frequency must be found and maintained.

5.2 Study Design

Sample Size. Low participant numbers are common across mood literature and were also a key trend within our sample. With 76% of studies having 0–100 participants, we aim to identify some driving factors behind this trend, as well as the challenges that arise when participant numbers grow. User dropouts, noncompliance, and periods of non-reporting are frequently referenced across the literature. Hence, by increasing the ability of these applications to process inconsistent and scarce data, raising participant numbers would aid in the wider adoption of smartphones as a mental health tool [85] while enhancing the validity of numerous studies. Indeed, this would introduce a series of challenges.

First, as participant numbers grow, methods to encourage sustained engagement become more limited [85]. By extension, non-compliance and dropout rates are likely to increase. Further, as most mood services require consistent data from self-reports and passive sensors, their effectiveness gets dampened when processing data that is noisy and inconsistent [80,85]. To address this challenge, existing study designs would need to be scaled to wider populations and be able to draw clinical conclusions from less consistent data. However, this would significantly enhance the difficulty of the overall process [80].

Despite these challenges, some studies have been successful using larger participant numbers. Servia-Rodríguez et al. [83] derived data from 18,000 participants for their study using a public Android application. They were still able to achieve day-level prediction accuracies of 61–63.5%, which is lower – but still comparable to similar work. Spathis et al. [85] ran a study using data from 17,251 participants over 3 years. They were able to reach day-level accuracies of 74%, which is a significant achievement given the large number of participants. These studies show that it is possible to achieve comparable prediction accuracy with significantly larger populations.

However, these studies also have the benefit of being able to pick users from their database who, longitudinally, have (1) completed a sufficient number of self-reports and (2) provided sufficient passive sensor data. While possible, it remains challenging to implement a clinically sound, large-scale mood-tracking study that does not employ data-cleaning techniques to remove noisy participant data from analysis. Jaques et al. [47] demonstrate how deep learning techniques can be used to account for missing sensor data. Using such techniques, this study was able to achieve strong mood prediction in situations where even up to 75% of data was lost. Future studies should consider implementing similar deep-learning techniques to deal with noisy or inconsistent data. This will facilitate larger participant samples, and henceforth extend mood prediction capabilities to a more realistic data environment where noise and inconsistency are more frequent.

Demography. Our results show that 13 of 28 applicable studies (46%) were comprised of undergraduates, postgraduates, and university staff. A further four studies recruited over 50% of their participants from universities. With most studies having 1–100 participants, and at least 43% of studies based on this homogeneous participant background, it is pungent to discuss whether some population groups are misrepresented in mood inference literature.

Universities are a convenient place to recruit participants, however, inferences gained from personalised student data are mostly ineffective in generalising to members of the general public [15,41]. For instance, Wang et al. [96] recruited 83 students over two 9-week terms, finding that their reliance on university routines eroded the potential for depression symptoms features to generalise to standard populations. Across the literature, some evidence is provided to explain this phenomenon.

First, relying exclusively on student data may create systematic biases as students generally have more dynamic attitudes, less formulated peer relationships, and stronger intellectual skills than the general population [15]. These systematic biases will inherently invalidate some findings. Second, students are also more likely to come from homogeneous backgrounds. A common criticism across psychology literature, therefore, is that claims are mostly based on data from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) cultures [43]. To the detriment of these studies, WEIRD societies rank very poorly in their ability to represent a general population [43]. When student data is personalised, such as in the case of mood inference studies, this effect is heightened [41].

On the basis of this evidence, validity concerns are raised for the 46% of studies that only used university participants. Future studies that intend on demonstrating findings that are applicable to general populations should be careful drawing conclusions from this narrow group of society [43]. To mitigate this, future studies should aim to increase the number of non-university participants.

5.3 Platform

Our analysis found that 20 of 36 studies (56%) used a service that was only accessible on Android. Conversely, 2 of 36 studies (6%) only offered their service on iOS. With 26 of 36 (72%) studies offering a service that is only accessible on one type of Smartphone OS, significant implications for accessibility are noted across the study set. As the two largest smartphone OS providers transcend into maturity, devices last longer, people are upgrading less and also becoming more content with their choices and are less likely to change their devices [5].

According to our results, Android's market share makes this less problematic for the 56% of Android-based studies, missing 29% of potential users highlights accessibility issues in this context. Yet, these concerns are far more severe for the 6% of studies that are iOS only. This group fails to reach 72% of potential users. Of the studies that mention the reason for their preference, Bachmann et al. [7] mentioned that Android was chosen as the exclusive offering due to having the

highest market share. Further, Cao et al. [25] used a specialised custom keyboard for mood detection, a feature available on Android that is restricted on iOS.

Indeed, most studies do not specify why they chose to be exclusive in their application offering. Some evidence in this domain suggests that it is difficult to develop for both Android and iPhone, due to differences in platform, tools, and techniques [92]. While this is likely to be a substantial driving factor towards many mood applications being exclusively offered on one operating system, further evidence must be considered to understand the clear preference for Android. More so than iOS, Android offers less restrictive permissions to access and collect user information and sensor data [1]. In the context of mood applications that track and record user data, this is a significant factor.

With Android having a significantly greater market share and flexibility in permissions, it is reasonable that 56% of studies were Android-only. The aforementioned studies [7] add further evidence to this claim. Significant gains in **accessibility** have been gained through a transition from pen-and-paper to the in-situ mobile domain; however, to maximise accessibility, adjustments can and should be made to offer corresponding services across all mobile platforms.

5.4 Privacy

As mood applications grow in efficacy and become more pervasive throughout society, privacy implications become more critical. Mood applications track and store significant amounts of idiosyncratic longitudinal data which could be used to gain, for instance, valuable personalised marketing insights [10]. This data could also be used for criminal intent, such as identity fraud. With increased internet usage throughout the pandemic, rates of cybercrime have grown significantly. Outcomes are worsened for people with mental illness as they have a reduced capacity to protect themselves online [68].

Health applications are generally run commercially and thus may not possess medically sound controls over sensitive data [46]. Further, the majority of health applications are not subject to government supervision, hence consumers and clinical advisers must provide their own technical inquiry into the privacy procedures of a service [45]. Hamre-Os [40] interviewed 26 students about mood applications and found that the potential for cyber-attacks created the greatest concern as this could lead to sensitive data being leaked to nefarious parties. Further, Widnall et al. [100] conducted a systematic analysis of mood application user reviews, finding that mistrust had risen after a series of high-profile cyber-attacks. By extension, some users felt uncomfortable because of a general lack of transparency with respect to data storage. In our results, 27 of 36 (75%) studies were transparent about attempting to preserve privacy in some manner.

However, these details were mostly brief and did not explicitly cover all bases. For instance, many studies covered their hashing method, but no studies went into explicit technical detail about how their server was secured. Hence, while also acknowledging that 10 studies did not report any privacy information whatsoever, it is understandable that some participants may feel like they are overextending their trust. To mitigate this, designers of mood services that store

personal data must communicate an explicit privacy strategy that covers data logging, storage, and server security.

5.5 Sensors

Regular intervals of information must be derived from passive sensors to facilitate the creation of predictive mood models. Not only does in-situ mood reporting give primary quantifiable mood data, but data is also provided in a context such that it can be correlated with real-life occurrences. Time-series data from passive sensors can be harnessed as a means of gaining information about these occurrences [55]. However, there are challenges regarding the heterogeneity of information, as sensors produce data at different rates and can give conflicting insights [106]. We, therefore, discuss the implications of sensor usage that were observed in the study set. Our results show that the most used sensors were the microphone and accelerometer, both appearing in 12 studies. This was followed by Application usage (10) and SMS/Call logs (9).

Microphone. Due to innovations in voice processing, the use and storage of microphone data have significant implications for privacy [54]. Using advanced ML techniques, speech can be analysed to make advanced inferences about a person, such as their personality type [73,74], gender, age, and socioeconomic background [54]. For this reason, most studies did not continuously record or store raw microphone data. In our sample, 12 studies used the microphone. The studies [6,34,83,85] used the microphone, but were explicit about only recording amplitude at various intervals throughout the day.

Zhang et al. [105] derived mean, variance, and other features from microphone audio before it was sent to the server, such that raw audio was never kept. Chang et al. [28] used the microphone as its only sensor but performed all computations locally to preserve privacy. Wang et al. [96] ran a conversation classifier continuously on each participant's phone which can detect speech segments and unique conversations. However, the classifier is unable to identify unique speakers, and speech data was uploaded over private Wi-Fi to a secure server.

While most studies provided a sound privacy-preserving method for microphone data, there was one exception. Lu et al. [62] extracted features such as mel-frequency cepstrum, speaking rate, and pitch range, all of which can be used to identify the speaker. This data was also stored externally for model training without providing any information about hashing or security. Overall, the usage of the microphone appears to be, for the most part, privacy-preserving. To mitigate privacy concerns, future developments in this field should continue to extract and store features from speech that can't be traced back to a specific human.

GPS. Similar to the microphone, the storage of GPS data is a key privacy concern as it can reveal enough sensitive information to identify users [37]. GPS data can be used to track a user's location, time spent there, and frequently visited sites, and to derive information about the user's everyday routines [24]. However, while

invasive, usage and storage of GPS data is generally an expected trade-off for most mood tracking applications due to the functional benefits it provides [10].

In our review, 8 papers used the GPS sensor. Most studies appear to have sound privacy processes, involving user consent, hashing, coarse-grained data collection and/or secure storage. However, most studies fail to cover all of these processes. For instance, LiKamWa et al. [61] hashed all private data but did not mention user consent nor provide a mechanism to opt out of location sharing. Other studies [69,96] sought approval from their Institutional Review Boards and stored GPS data onto secure servers when the users' phones were connected to a private Wi-Fi network. However, they did not mention that this data was hashed or encrypted.

Servia-Rodríguez et al. [83] mitigated their GPS data collection by predominantly using Wi-Fi and Cell towers to derive location. When they did collect location data, it was correlated with a self-reported location such that only coarse-grained location was recorded. While this strategy is highly effective, this study did not mention that this data was hashed, nor did it state that the upload server was secured. This is potentially problematic, as it is foreseeable that even coarse-grained location data could have severe privacy implications if leaked.

Canzian and Musolesi [24] was explicit about receiving approval from the Ethics Review Board, provided a consent form, and uploaded GPS data via a 'secure transmission protocol' to a secure server. While user data was not hashed or encrypted, this can be considered a reasonable privacy strategy due to other robust mechanisms. Conversely, Lu et al. [62] collected GPS data every three minutes but provided no information about hashing, obscuration, server security or the ability to opt out. To address GPS privacy concerns in a more holistic manner, future studies should (1) incorporate consent forms and the ability to opt out, (2) hash or encrypt all data, (3) only collect coarse-grained location, and (4) store data securely into a secure server.

5.6 Smartphone as 'ubiquitous instrumentation'

As smartphones are equipped with a variety of sensors and closely accompany their owners throughout their daily lives, smartphones can serve as a powerful tool for unobtrusive and continuous mood tracking. Smartphone sensors that track our physical activities can provide detailed insights into users' emotional states and moods based on activity trackers' data [16]. Furthermore, prior research has shown that application usage can be leveraged as a reliable estimate to predict users' emotional states and mood [81]. Finally, textual information posted on social media platforms, as well as other written communication, can be indicative of a user's emotional states and mood, with certain words being associated with either positive or negative affect.

Due to the rise in the sensing and computing abilities of smartphones, these devices are well-suited for observing human behaviour in ways that current scientific methods are unable to do. Particularly given that smartphone ownership is accessible to large parts of the population, in contrast to traditional scientific equipment that typically requires significant financial investment from governments or institutions, making them perfect for 'ubiquitous instrumentation' [52].

Our results reflect the prominent role of smartphones in emotion tracking and mood inference in field research and laboratory-based user studies. Our findings show that sensors such as accelerometers, gyroscopes, microphones, and cameras can detect changes in movement, tone of voice, and facial expressions, which can indicate the user's emotional state. Researchers can analyse these signals and identify patterns to infer the user's emotional state without invading user privacy. Moreover, our results show that smartphones provide an excellent platform for collecting self-reported data from users about their moods and emotional states. While self-reported data may be subject to biases, it can provide valuable context when combined with sensor data [103].

5.7 Implications for HCI Research

When designing and conducting studies using smartphones for emotion sensing and mood tracking, several aspects of these devices must be considered. To be precise, careful sensor collection for mood tracking is crucial. Sensors vary in accuracy and invasiveness, thus, the ethical implications of using privacy-comprising sensors like microphones and cameras must be taken into account.

To address privacy concerns and protect private data, researchers should consider hashing or encrypting the data and securely hosting it. Obtaining informed consent from participants, and clearly explaining data collection and usage, is essential. Measures should be implemented to safeguard participant data against unauthorised access or misuse. Additionally, potential sensor data loss or noise should be acknowledged, and deep-learning techniques can be employed to handle inconsistent or noisy data effectively.

Furthermore, the use of additional sensing devices alongside smartphones in user studies should be carefully considered. Our findings indicate that the majority of existing work in the field can be conducted using smartphones alone. Therefore, it is recommended to avoid additional instrumentation unless absolutely necessary to ensure technology accessibility in user studies.

By considering factors such as prevalent sensor usage, study samples, privacy concerns, and accessibility, HCI researchers can make informed decisions when designing user studies, leading to more accurate and meaningful outcomes.

6 Conclusion

In this paper, we present the review and analysis of the ACM body of literature surrounding mood tracking and inference using mobile devices. We differentiated mood from emotion, before explaining the motivation and reasoning to measure mood with respect to clinical disorders and personal empowerment. We then looked into how mood can be tracked and predicted using clinical tools, including how smartphones have significantly increased the efficacy of such tools. We analysed 32 papers from the Association for Computing Machinery (ACM) Digital Library and discussed mood reporting strategy, sensors, study design, demography, platforms, and privacy. Based on our findings, we develop a set of recommendations concerning self-reporting methods, sensor choices, platform, accessibility, and privacy considerations.

A Appendix

Table 1. Frequency of daily mood input and number of studies ($N = 24$).

Frequency of Daily Mood Inputs	Num Studies
1x	6
2x	8
3x	4
4x	4
12x	2

Table 2. Summary of Metrics selected and filled in per paper.

Paper	Mic	GPS	Accele-rometer	Application Usage	SMS/Call Info	Calendar
[61]	No	Yes	No	Yes	Yes	No
[25]	No	No	Yes	No	No	No
[86]	No	No	No	No	No	No
[105]	Yes	Yes	Yes	No	Yes	No
[6]	Yes	No	No	Yes	Yes	Yes
[28] [1/2]	Yes	No	No	No	No	No
[28] [2/2]	Yes	No	Yes	No	No	No
[96]	Yes	Yes	Yes	Yes	Yes	No
[106]	Yes	No	Yes	Yes	No	No
[62]	Yes	Yes	Yes	No	No	No
[17]	No	No	No	No	Yes	No
[69]	Yes	Yes	Yes	Yes	Yes	No
[56]	No	No	No	No	No	No
[39]	No	No	No	No	No	No
[80] [1/2]	No	No	No	No	No	No
[80] [2/2]	No	No	No	No	No	No
[34]	Yes	No	No	Yes	Yes	Yes
[83]	Yes	Yes	Yes	No	Yes	No
[31]	No	Yes	Yes	No	Yes	No
[67]	No	No	No	No	No	No
[24]	No	Yes	No	No	No	No
[85]	Yes	No	Yes	No	No	No
[42] *	Yes *	Yes *	Yes	No	No	No
[91]	-	-	-	-	-	-
[94]	-	-	-	-	-	-
[51]	-	-	-	-	-	-
[93] [1/2]	No	No	No	Yes	No	No
[93] [2/2]	No	No	No	Yes	No	No
[60]	-	-	-	-	-	-
[59]	No	No	Yes	No	No	No
[7] *	Yes	No	Yes	Yes	Yes	Yes
[29]	-	-	-	-	-	-
[70]	-	-	-	-	-	-
[104]	No	No	No	Yes	No	No
[8]	-	-	-	-	-	-
[2]	Yes	No	Yes	Yes	No	No
[58]	No	No	Yes	No	No	No
[18]	-	-	-	-	-	-
[64] [1/2]	-	-	-	-	-	-
[64] [2/2]	-	-	-	-	-	-

Table 3. Summary of Metrics selected and filled in per paper.

Paper	Light Sensor	Screen On/Off	WiFi	Bluetooth	Physical Activity	Keyboard
citech22likamwa2013	No	No	No	No	No	No
[25]	No	No	No	No	No	Yes
[86]	No	No	No	No	No	No
[105]	Yes	Yes	Yes	No	Yes	No
[6]	Yes	No	Yes	No	Yes	No
[28]	No	No	No	No	No	No
[28] Study 2	No	No	No	No	No	No
[96]	Yes	Yes	Yes	Yes	Yes	No
[106]	Yes	Yes	Yes	No	Yes	No
[62]	No	No	No	No	Yes	No
[17]	No	No	No	Yes	No	No
[69]	Yes	Yes	Yes	Yes	Yes	No
[56]	No	No	No	No	No	No
[39]	No	No	No	No	No	No
[80]	No	No	No	No	No	No
[80] Study 2	No	No	No	No	No	No
[34]	Yes	No	Yes	No	Yes	No
[83]	No	No	Yes	No	Yes	No
[31]	No	No	Yes	No	Yes	No
[67]	No	No	No	No	No	No
[24]	No	No	No	No	Yes	No
[85]	No	No	No	No	Yes	No
[42] *	No	No	No	No	Yes	No
[91]	-	-	-	-	-	-
[94]	-	-	-	-	-	-
[51]	-	-	-	-	-	-
[93]	No	No	No	No	No	No
[93] Study 2	No	No	No	No	No	No
[60]	-	-	-	-	-	-
[59]	No	No	No	No	Yes	No
[7] *	Yes	No	Yes	No	No	No
[29]	-	-	-	-	-	-
[70]	-	-	-	-	-	-
[104]	No	No	No	No	No	Yes
[8]	-	-	-	-	-	-
[2]	No	Yes	Yes	Yes	No	No
[58]	No	No	No	No	No	No
[18]	-	-	-	-	-	-
[64]	-	-	-	-	-	-
[64] Study 2	-	-	-	-	-	-

Table 4. Summary of Metrics selected and filled in per paper.

Paper	In-situ self reporting?	Self reporting for ground truth?	Passive tracking?
[61]	Yes	Yes	Yes
[25]	No	No	Yes
[86]	Yes	No	No
[105]	Yes	Yes	Yes
[6]	Yes	Yes	Yes
[28]	No	No	Yes
[28] Study 2	No	No	Yes
[96]	Yes	Yes	Yes
[106]	Yes	Yes	Yes
[62]	No	No	Yes
[17]	Yes	Yes	Yes
[69]	Yes	Yes	Yes
[56]	Yes	Yes	Yes
[39]	Yes	No	No
[80]	Yes	Yes	Yes
[80] Study 2	Yes	Yes	Yes
[34]	Yes	Yes	Yes
[83]	Yes	Yes	Yes
[31]	No	No	Yes
[67]	No	No	Yes
[24]	Yes	Yes	Yes
[85]	Yes	Yes	Yes
[42] *	Yes	Yes	Yes
[91]	Yes	No	No
[94]	N/A	N/A	N/A
[51]	Yes	No	No
[93]	Yes	Yes	Yes
[93] Study 2	Yes	Yes	Yes
[60]	N/A	N/A	N/A
[59]	Yes	Yes	Yes
[7] *	No	No	Yes
[29]	Yes	No	No
[70]	Yes	No	No
[104]	Yes	Yes	Yes
[8]	Yes	No	No
[2]	Yes	Yes	Yes
[58]	Yes	Yes	Yes
[18]	Yes	No	No
[64]	Yes	No	No
[64] Study 2	Yes	No	No

Table 5. Summary of Metrics selected and filled in per paper.

Paper	States privacy process?	Length of self-reporting period	Daily active mood input frequency	Android or iOS
[61]	Yes	2 months	4x	Both
[25]	Yes	N/A	N/A	Android
[86]	Yes	2 weeks	3x	Both
[105]	Yes	1 month	3x	Android
[6]	Yes	4 days	12x	Android
[28]	Yes	N/A	N/A	All
[28] Study 2	Yes	N/A	N/A	All
[96]	Yes	18 weeks	4x	Both
[106]	No	2 weeks	3x	Android
[62]	No	N/A	N/A	Android
[17]	Yes	2 weeks	1x	Android
[69]	Yes	3 weeks	4x	Both
[56]	Yes	28 days	1x	Both
[39]	Yes	-	1x	iOS
[80]	Yes	5 weeks	4x	Android
[80] Study 2	Yes	-	-	Android
[34]	Yes	1 month	12x	Android
[83]	Yes	26 days	2x	Android
[31]	No	N/A	N/A	-
[67]	Yes	N/A	N/A	-
[24]	Yes	-	1x	Android
[85]	Yes	-	2x	Android
[42] *	Yes	-	-	iOS
[91]	Yes	N/A	2x	Android
[94]	N/A	N/A	N/A	N/A
[51]	Yes	-	2x	Android
[93]	Yes	2 weeks	2x	Android
[93] Study 2	Yes	2 weeks	2x	Android
[60]	N/A	N/A	N/A	N/A
[59]	No	-	-	N/A
[7] *	No	-	-	Android
[29]	No	N/A	N/A	iOS
[70]	No	N/A	-	Both
[104]	Yes	-	2x	Android
[8]	N/A	N/A	-	-
[2]	Yes	N/A	1x	Android
[58]	Yes	N/A	1x	N/A
[18]	Yes	N/A	3x	Both
[64]	Yes	N/A	-	Android
[64] Study 2	Yes	N/A	-	Android

Table 6. Summary of Metrics selected and filled in per paper.

Paper	Multiple devices?	Study details	*N*	In-the-wild/ Lab	Duration	Participant background	Gender Split (% female)
[61]	No	*	32	wild	2 months	24/32 University	34.4
[25]	No	*	21	wild	8 weeks	-	-
[86]	No	*	2382	wild	22 months	Random mixed	-
[105]	No	*	42	wild	1 month	University	57
[6]	No	*	9	wild	4 days	-	44
[28]	No	*	125	lab	N/A	-	75
[28] Study 2	No	*	7	lab	N/A	-	43
[96]	Yes	*	83	wild	18 weeks	University	52
[106]	No	*	68	wild	6 weeks	-	-
[62]	Yes	*	14	lab	4 days	University	71
[17]	No	*	111	wild	7 months	University	-
[69]	Yes	*	83	wild	18 weeks	University	52
[56]	No	*	36	wild	28 days	University	-
[39]	No	*	9	lab	1 day	University	33
[80]	No	*	23	wild	5 weeks	University	40
[80] Study 2	No	*	9654	wild	-	Random mixed	N/A
[34]	Yes	*	6	wild	1 month	4 University, 2 Workers	33
[83]	No	*	18000	wild	35 months	Random mixed	N/A
[31]	No	*	100	wild	1 month	-	-
[67]	No	*	100	lab	-	Random mixed	N/A
[24]	No	*	28	wild	9 months	University	54
[85]	No	*	17251	wild	46 months	Random mixed	N/A
[42] *	Yes	N/A	N/A	N/A	N/A	N/A	N/A
[91]	No	*	391	wild	2 weeks	Random mixed	N/A
[94]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[51]	No	*	48	wild	11 days	University	40
[93]	No	*	15	wild	2 weeks	University	50
[93] Study 2	No	*	21	wild	2 weeks	University	50
[60]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
[59]	No	*	2	-	-	-	-
[7] *	Yes	N/A	N/A	N/A	N/A	N/A	N/A
[29]	No	*	15	wild	2 weeks	Professionals, University	27
[70]	No	*	17	wild	2 weeks	-	-
[104]	No	*	30	wild	1 year	University	-
[8]	Yes	*	22	lab	1 day	School teenagers	45
[2]	No	*	18	wild	5 months	Mixed bipolar patients	-
[58]	No	*	255	wild	2 months	University	-
[18]	No	*	1461	wild	15 months	Random mixed	-
[64]	No	*	6	wild	1 day	School teenagers	-
[64] Study 2	No	*	73	wild	2 weeks	School teenagers	86

References

1. Alshehri, A., Hewins, A., McCulley, M., Alshahrani, H., Fu, H., Zhu, Y.: Risks behind device information permissions in android OS. Commun. Netw. **09**(04), 219–234 (2017)

2. Alvarez-Lozano, J., et al.: Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state. In: PETRA 2014 (2014)
3. Ang, W.T., Khosla, P.K., Riviere, C.N.: Nonlinear regression model of alow-g mems accelerometer. IEEE Sens. J. **7**, 81–88 (2007)
4. Angst, J., Cassano, G.: The mood spectrum: improving the diagnosis of bipolar disorder. Bipolar Disord. **7**(s4), 4–12 (2005)
5. Appiah, D., Ozuem, W., Howell, K.: Brand switching in the smartphone industry: a preliminary study (2017)
6. Bachmann, A., et al.: How to use smartphones for less obtrusive ambulatory mood assessment and mood recognition. In: UbiComp/ISWC 2015 Adjunct, pp. 693–702 (2015)
7. Bachmann, A., et al.: Leveraging smartwatches for unobtrusive mobile ambulatory mood assessment. In: UbiComp/ISWC 2015 Adjunct, pp. 1057–1062 (2015)
8. Balta, A., Read, J.C.: U ok? Txt me the colour of ur mood! In: CHI EA 2016, pp. 2410–2416 (2016)
9. Bankmycell: How many smartphones are in the world? (2022). https://www.bankmycell.com/blog/how-many-phones-are-in-the-world
10. Barcena, M.B., Wueest, C., Lau, H.: How safe is your quantified self? Technical report, Symantec, Mountain View, CA (2014)
11. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol. Sci. Public Interest **20**, 1–68 (2019)
12. Beedie, C., Terry, P., Lane, A.: Distinctions between emotion and mood. Cogn. Emot. **19**(6), 847–878 (2005)
13. van Berkel, N., Ferreira, D., Kostakos, V.: The experience sampling method on mobile devices. ACM Comput. Surv. **50**(6) (2017)
14. van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., Kostakos, V.: Overcoming compliance bias in self-report studies: a cross-study analysis. Int. J. Hum. Comput. Stud. **134**, 1–12 (2020)
15. van Berkel, N., Sarsenbayeva, Z., Goncalves, J.: The methodology of studying fairness perceptions in artificial intelligence: contrasting chi and FAccT. Int. J. Hum. Comput. Stud. **170**, 102954 (2023)
16. Biddle, S.J.H.: Emotion, mood and physical activity, pp. 75–97 (2003)
17. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Pentland, A.S.: Daily stress recognition from mobile phone data, weather conditions and individual traits. In: MM 2014, pp. 477–486 (2014)
18. Bond, R., Moorhead, A., Mulvenna, M., O'Neill, S., Potts, C., Murphy, N.: Behaviour analytics of users completing ecological momentary assessments in the form of mental health scales and mood logs on a smartphone app. In: ECCE 2019, pp. 203–206 (2019)
19. Bower, G.H.: Mood and memory. Am. Psychol. **36**(2), 129–148 (1981)
20. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)
21. Brown, T.A., Chorpita, B.F., Barlow, D.H.: Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. J. Abnorm. Psychol. **107**(2), 179–192 (1998)
22. Caldeira, C.M., Chen, Y., Chan, L., Pham, V., Chen, Y., Zheng, K.: Mobile apps for mood tracking: an analysis of features and user reviews. In: AMIA ... Annual Symposium Proceedings. AMIA Symposium 2017, pp. 495–504 (2017)

23. Calear, A., Christensen, H.: Systematic review of school-based prevention and early intervention programs for depression. J. Adolesc. **33**, 429–438 (2009)
24. Canzian, L., Musolesi, M.: Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In: UbiComp 2015, pp. 1293–1304 (2015)
25. Cao, B., et al.: DeepMood: modeling mobile phone typing dynamics for mood detection. In: KDD 2017, pp. 747–755 (2017)
26. Carlson, M., Charlin, V., Miller, N.: Positive mood and helping behavior: a test of six hypotheses. J. Pers. Soc. Psychol. **55**(2), 211–229 (1988)
27. Chan, E.C., Sun, Y., Aitchison, K.J., Sivapalan, S.: Mobile app–based self-report questionnaires for the assessment and monitoring of bipolar disorder: systematic review. JMIR Formative Res. **5**(1), e13770 (2021)
28. Chang, K.H., Fisher, D., Canny, J., Hartmann, B.: How's my mood and stress? An efficient speech analysis library for unobtrusive monitoring on mobile phones. In: BodyNets 2011, pp. 71–77. ICST (2011)
29. Church, K., Hoggan, E., Oliver, N.: A study of mobile mood awareness and communication through MobiMood. In: NordiCHI 2010, pp. 128–137 (2010)
30. Depp, C., Kim, D., Dios, L., Wang, V., Ceglowski, J.: A pilot study of mood ratings captured by mobile phone versus paper- and-pencil mood charts in bipolar disorder. J. Dual Diagn. **8**, 326–332 (2012)
31. Dhahri, C., Ikeda, K., Hoashi, K.: Forecasting mood using smartphone and SNS data. In: HotMobile 2019, p. 175 (2019)
32. Diener, E., Wirtz, D., Tov, W.: New measures of well-being: flourishing and positive and negative feelings. Soc. Indic. Res. **39**, 247–266 (2010)
33. Dubad, M., Winsper, C., Meyer, C., Livanou, M., Marwaha, S.: A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. Psychol. Med. **48**, 1–21 (2017)
34. Exler, A., Schankin, A., Klebsattel, C., Beigl, M.: A wearable system for mood assessment considering smartphone features and data from mobile ECGs. In: UbiComp 2016, Adjunct, pp. 1153–1161 (2016)
35. Gardner, M.P.: Mood states and consumer behavior: a critical review. J. Consum. Res. **12**(3), 281 (1985)
36. George, J.M.: State or trait: effects of positive mood on prosocial behaviors at work. J. Appl. Psychol. **76**(2), 299–307 (1991)
37. Goldenholz, D.M., et al.: Using mobile location data in biomedical research while preserving privacy. J. Am. Med. Inform. Assoc. **25**(10), 1402–1406 (2018)
38. Gross, J.J.: Emotion regulation: current status and future prospects. Psychol. Inq. **26**(1), 1–26 (2015)
39. Hafiz, P., Maharjan, R., Kumar, D.: Usability of a mood assessment smartphone prototype based on humor appreciation. In: MobileHCI 2018, Adjunct, pp. 151–157 (2018)
40. Hamre-Os, A.: A mood tracking interface for mobile application-to help assess well being in students (2021)
41. Hanel, P.H.P., Vione, K.C.: Do student samples provide an accurate estimate of the general public? PLoS ONE **11**(12), e0168354 (2016)
42. Hänsel, K., Alomainy, A., Haddadi, H.: Large scale mood and stress self-assessments on a smartwatch. In: UbiComp 2016, Adjunct, pp. 1180–1184 (2016)
43. Henrich, J., Heine, S.J., Norenzayan, A.: The weirdest people in the world? Behav. Brain Sci. **33**(2–3), 61–83 (2010)

44. Hibbard, J.H., Stockard, J., Mahoney, E.R., Tusler, M.: Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv. Res. **39**(4p1), 1005–1026 (2004)
45. Huckvale, K., Torous, J., Larsen, M.: Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw. Open **2**, e192542 (2019)
46. Hutton, L., et al.: Assessing the privacy of mHealth apps for self-tracking: heuristic evaluation approach. JMIR Mhealth Uhealth **6**(10), e185 (2018)
47. Jaques, N., Taylor, S., Sano, A., Picard, R.: Multimodal autoencoder: a deep learning approach to filling in missing sensor data and enabling better mood prediction, pp. 202–208 (2017)
48. Kessler, R.C., Berglund, P.A., Foster, C.L., Saunders, W.B., Stang, P.E., Walters, E.E.: Social consequences of psychiatric disorders, II: teenage parenthood. Am. J. Psychiatry **154**(10), 1405–1411 (1997)
49. Kessler, R.C., Foster, C.L., Saunders, W.B., Stang, P.E.: Social consequences of psychiatric disorders, i: educational attainment. Am. J. Psychiatry **152**(7), 1026–1032 (1995)
50. Kessler, R.C., Walters, E.E., Forthofer, M.S.: The social consequences of psychiatric disorders, III: probability of marital stability. Am. J. Psychiatry **155**(8), 1092–1096 (1998)
51. Khue, L.M., Ouh, E.L., Jarzabek, S.: Mood self-assessment on smartphones. In: WH 2015 (2015)
52. Kostakos, V., Ferreira, D.: The rise of ubiquitous instrumentation. Frontiers ICT **2**, 3 (2015)
53. Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The PHQ-9. J. Gen. Intern. Med. **16**(9), 606–613 (2001)
54. Kröger, J.L., Lutz, O.H.-M., Raschke, P.: Privacy implications of voice and speech analysis – information disclosure by inference. In: Friedewald, M., Önen, M., Lievens, E., Krenn, S., Fricker, S. (eds.) Privacy and Identity 2019. IAICT, vol. 576, pp. 242–258. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-42504-3_16
55. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Comm. Mag. **48**(9), 140–150 (2010)
56. Lee, J.A., Efstratiou, C., Bai, L.: OSN mood tracking: exploring the use of online social network activity as an indicator of mood changes. In: UbiComp 2016, Adjunct, pp. 1171–1179 (2016)
57. Lee, K., et al.: Effect of self-monitoring on long-term patient engagement with mobile health applications. PLoS ONE **13**, e0201166 (2018)
58. Li, B., Sano, A.: Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **4**(2) (2020)
59. Lietz, R., Harraghy, M., Brady, J., Calderon, D., Cloud, J., Makedon, F.: A wearable system for unobtrusive mood detection. In: PETRA 2019, pp. 329–330 (2019)
60. Lietz, R., Harraghy, M., Calderon, D., Brady, J., Becker, E., Makedon, F.: Survey of mood detection through various input modes. In: PETRA 2019, pp. 28–31 (2019)
61. LiKamWa, R., Liu, Y., Lane, N.D., Zhong, L.: MoodScope: building a mood sensor from smartphone usage patterns. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys 2013, pp. 465–466 (2013)

62. Lu, H., et al.: StressSense: detecting stress in unconstrained acoustic environments using smartphones. In: UbiComp 2012, pp. 351–360 (2012)
63. Lyubomirsky, S., King, L., Diener, E.: The benefits of frequent positive affect: does happiness lead to success? Psychol. Bull. **131**(6), 803–855 (2005)
64. Matthews, M., Doherty, G.: In the mood: engaging teenagers in psychotherapy using mobile phones. In: CHI 2011, pp. 2947–2956 (2011)
65. Mehrotra, A., Vermeulen, J., Pejovic, V., Musolesi, M.: Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling (2015)
66. Mitchell, R.L., Phillips, L.H.: The psychological, neurochemical and functional neuroanatomical mediators of the effects of positive and negative mood on executive functions. Neuropsychologia **45**(4), 617–629 (2007)
67. Mogadala, A., Varma, V.: Twitter user behavior understanding with mood transition prediction. In: DUBMMSM 2012, pp. 31–34 (2012)
68. Monteith, S., Bauer, M., Alda, M., Geddes, J., Whybrow, P.C., Glenn, T.: Increasing cybercrime since the pandemic: concerns for psychiatry. Current Psychiatry Rep. **23**(4) (2021)
69. Morshed, M.B., et al.: Prediction of mood instability with passive sensing. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(3) (2019)
70. Nolasco, H.R., Waldman, M., Vargo, A.W.: Exploring emotional reappraisal and repression through acoustic mood self-tracking. In: UbiComp 2021, Adjunct, pp. 248–252 (2021)
71. Patel, R., et al.: Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. BMJ Open **5**(5), e007504–e007504 (2015)
72. Penedo, F.J., Dahn, J.R.: Exercise and well-being: a review of mental and physical health benefits associated with physical activity. Curr. Opin. Psychiatry **18**(2), 189–193 (2005)
73. Polzehl, T.: Personality in Speech. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09516-5
74. Polzehl, T., Möller, S., Metze, F.: Automatically assessing acoustic manifestations of personality in speech. In: 2010 IEEE Spoken Language Technology Workshop, pp. 7–12 (2010)
75. Rickwood, D., Deane, F.P., Wilson, C.J., Ciarrochi, J.: Young people's help-seeking for mental health problems. Aust. e-J. Adv. Mental health **4**(3), 218–251 (2005)
76. Rideout, V., Fox, S., Peebles, A., Robb, M.B.: Coping with Covid-19: how young people use digital media to manage their mental health. Common Sense and Hopelab, San Francisco, CA (2021)
77. Rottenberg, J.: Mood and emotion in major depression. Curr. Dir. Psychol. Sci. **14**(3), 167–170 (2005)
78. Russell: Core affect and the psychological construction of emotion. Psychol. Rev. **110**(1), 145–172 (2003)
79. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
80. Saha, K., Chan, L., De Barbaro, K., Abowd, G.D., De Choudhury, M.: Inferring mood instability on social media by leveraging ecological momentary assessments. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **1**(3) (2017)
81. Sarsenbayeva, Z., et al.: Does smartphone use drive our emotions or vice versa? A causal analysis. In: CHI 2019, pp. 1–15 (2020)

82. Schueller, S., Neary, M., Lai, J., Epstein, D.: Understanding people's use of and perspectives on mood tracking apps: an interview study (preprint). JMIR Mental Health **8** (2021)
83. Servia-Rodríguez, S., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., Lathia, N., Sandstrom, G.M.: Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In: WWW 2017, pp. 103–112 (2017)
84. Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. Annu. Rev. Clin. Psychol. **4**(1), 1–32 (2008)
85. Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., Rentfrow, J.: Passive mobile sensing and psychological traits for large scale mood prediction. In: PervasiveHealth 2019, pp. 272–281 (2019)
86. Suhara, Y., Xu, Y., Pentland, A.S.: DeepMood: forecasting depressed mood based on self-reported histories via recurrent neural networks. In: WWW 2017, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 715–724 (2017)
87. Tag, B., Goncalves, J., Webber, S., Koval, P., Kostakos, V.: A retrospective and a look forward: lessons learned from researching emotions in-the-wild. IEEE Pervasive Comput. **21**, 28–36 (2022)
88. Tag, B., Sarsenbayeva, Z., Cox, A.L., Wadley, G., Goncalves, J., Kostakos, V.: Emotion trajectories in smartphone use: towards recognizing emotion regulation in-the-wild. Int. J. Hum. Comput. Stud. **166**, 102872 (2022)
89. Tag, B., et al.: Making sense of emotion-sensing: workshop on quantifying human emotions. In: UbiComp/ISWC 2021 Adjunct, pp. 226–229 (2021)
90. Toegel, G., Anand, N., Kilduff, M.: Emotion helpers: the role of high positive affectivity and high self-monitoring managers. Pers. Psychol. **60**(2), 337–365 (2007)
91. Torkamaan, H., Ziegler, J.: Mobile mood tracking: an investigation of concise and adaptive measurement instruments. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **4**(4) (2020)
92. Tracy, K.: Mobile application development experiences on apple's iOS and android OS. IEEE Potentials **31**, 30–34 (2012)
93. Visuri, A., Sarsenbayeva, Z., Goncalves, J., Karapanos, E., Jones, S.: Impact of mood changes on application selection. In: UbiComp 2016, Adjunct, pp. 535–540 (2016)
94. Wallbaum, T., Heuten, W., Boll, S.: Comparison of in-situ mood input methods on mobile devices. In: MUM 2016, pp. 123–127 (2016)
95. Wang, P.S., et al.: Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes. JAMA **298**(12), 1401 (2007)
96. Wang, R., et al.: Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **2**(1) (2018)
97. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Soc. Psychol. **54**(6), 1063 (1988)
98. van der Watt, A.S.J., Odendaal, W., Louw, K., Seedat, S.: Distant mood monitoring for depressive and bipolar disorders: a systematic review. BMC Psychiatry **20**(1) (2020)
99. Wen, C.K.F., Schneider, S., Stone, A.A., Spruijt-Metz, D.: Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J. Med. Internet Res. **19**(4), e132 (2017)

100. Widnall, E., et al.: A qualitative content analysis of user perspectives of mood-monitoring apps available to young people. (preprint). JMIR mHealth and uHealth **8** (2020)
101. Yang, C.C., Hsu, Y.L.: A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors **10**(8), 7772–7788 (2010)
102. Yang, K., et al.: Survey on emotion sensing using mobile devices. IEEE Trans. Affect. Comput. (2022)
103. Yang, K., et al.: Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. IEEE Trans. Affect. Comput. **3045**, 1 (2021)
104. Zhang, H., Gashi, S., Kimm, H., Hanci, E., Matthews, O.: MoodBook: an application for continuous monitoring of social media usage and mood. In: UbiComp 2018, pp. 1150–1155 (2018)
105. Zhang, X., Li, W., Chen, X., Lu, S.: MoodExplorer: towards compound emotion detection via smartphone sensing. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **1**(4) (2018)
106. Zhang, X., Zhuang, F., Li, W., Ying, H., Xiong, H., Lu, S.: Inferring mood instability via smartphone sensing: a multi-view learning approach. In: MM 2019, pp. 1401–1409 (2019)