"I spent 14 hours debugging just one assignment": Toward Computer-Mediated Personal Informatics for Computer Science Student Mental Health

Aishwarya Chandrasekaran
University of Delaware
Newark, Delaware, USA
aishc@udel.edu

London Bielicke
Rhodes College
Memphis, Tennessee, USA
bielm-24@rhodes.edu

Diya Shah
University of Delaware
Newark, Delaware, USA
diyashah@udel.edu

Harisha Janakiraman
University of Delaware
Newark, Delaware, USA
harishaj@udel.edu

Matthew Louis Mauriello
University of Delaware
Newark, Delaware, USA
mlm@udel.edu

Abstract

Anxiety and depression rates in Computer Science (CS) students are double those of other undergraduates and 5-10 times higher than the general population. However, factors contributing to the elevated mental health issues in CS students remain unknown. To bridge this gap, we conducted need-finding interviews (N=20), which revealed that the complexity of debugging, along with imposter syndrome, are key contributors to stress and burnout. Participants expressed openness toward and feature preferences in a computer-based Personal Informatics (PI) tool to facilitate self-reflection. In response, we developed EmotionStream, an algorithm-assisted PI tool that provides both contextual and emotional insights based on individual behaviors. We found that participants rated their experience with the tool highly. Post-hoc analysis revealed that emotional states, augmented with contextual cues, show promise of predicting real-time stress. Based on our findings, we provide design implications for future PI tools to support CS student mental well-being.

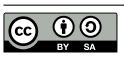
CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → *Health informatics.*

Keywords

student, stress, affect, personal informatics, computer science, mental health

ACM Reference Format:
Aishwarya Chandrasekaran, London Bielicke, Diya Shah, Harisha Janakiraman, and Matthew Louis Mauriello. 2025. "I spent 14 hours debugging just one assignment": Toward Computer-Mediated Personal Informatics for Computer Science Student Mental Health. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713269



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713269

1 Introduction

Young adults (ages 18-24) demonstrate the highest rate (39%) of mental illness [78], meaning many of their initial and ongoing experiences with psycho-emotional distress occur during college years [84]. A nationwide survey conducted in the U.S. in 2020 encompassing 30,725 undergraduate and 15,346 graduate and professional students reported major depressive and anxiety disorders in over a third of the population [26], demonstrating a trend toward escalating levels of severe mental health problems on campuses nationwide [40, 78, 87]. Specifically, a study conducted among U.S. engineering students revealed that they are nearly ten times more likely to exhibit a high risk of serious mental health disorders compared to the general U.S. adult population [34]. A similar study reported that the prevalence of anxiety and depression symptoms was twice as high in CS students compared to other undergraduate students and 5-10x higher than in the general population [103]. Notably, students in Computing fields face the highest risk of mental health disorders among all engineering disciplines [34]. Although recent research has explored the ways in which CS students may be able to reduce symptoms of anxiety and depression [37, 103], the reasons why CS students experience elevated mental health issues are unknown. Ji et al. have attempted to address this problem by investigating the reasons for stress among CS students, but the results are constrained only to students with pre-existing mental health conditions [52]. The specific academic challenges that contribute to elevated stress and burnout in the broader CS student population remain unclear.

In this paper, we investigate the unique challenges faced by CS students, particularly in relation to their academic work. Through needfinding interviews with 20 CS students, we identified debugging code, lack of self-awareness of stress, and imposter syndrome as the most stressful challenges stemming from academic work. Unsuccessful debugging attempts caused participants to get trapped in a cycle of error detection and the inadvertent introduction of new errors. Furthermore, participants reported that prolonged efforts to fix their code led to a lack of self-awareness of their stress and emotions, resulting in increased feelings of burnout and cycles of sleep deprivation. When asked about their willingness to adopt a computer-based PI tool, most participants indicated that they were open to using such a tool to facilitate self-reflection. They emphasized the need for real-time continuous emotional state monitoring

and stress reflection across various situational contexts (activity type, application type, and the privacy sensitivity of the activity), as well as within different temporal contexts.

To bridge this gap, we designed EmotionStream - an algorithm-assisted PI tool that provides both contextual and emotional insights based on individual behaviors. EmotionStream adopts a hybrid approach that combines algorithmic output and self-reports to promote self-reflection. EmotionStream has the following components: i) passive logging of emotional states through state-of-the-art Automated Emotion Recognition (AER) algorithms (DeepFace [97–99] and Residual Masking Network [85]) using facial cues, peripheral data (keystroke and mouse interactions), and application type ii) periodic collection of self-reported stress and emotional levels, stressors, activity type, time, and privacy sensitivity through Experience Sampling Methods (ESMs) [64], and iii) visualization to depict temporal trends of emotional states across situational contexts to increase students' awareness of their emotional responses to academic tasks. Finally, EmotionStream also allows participants to rate the system and provide feedback on the visualization dashboard.

To determine the acceptability of EmotionStream and validate AER for stress prediction, we conducted a week-long naturalistic and unconstrained study with 12 CS students. Results showed that participants' stress and emotional states varied with situational and temporal contexts. Specifically, higher stress values were reported during debugging and later parts of the night, corroborating our qualitative findings. Statistical analyses indicated a significant correlation between situational contexts and stress levels among CS students. Furthermore, tool ratings and engagement demonstrated the acceptability of EmotionStream. Moreover, the alignment between AER model classifications and self-reported emotional states yielded accuracies of 42% and 47% for Residual Masking Network (RMN) and DeepFace, respectively. Secondary analysis to predict momentary stress yielded an F1 score of 0.88 when augmenting contextual features with self-reported emotional states, highlighting the role of context in mental health. Our contributions include:

- Qualitative insights regarding the unique challenges of CS students, specifically due to academic tasks and their preferences for technological support.
- A novel computer-mediated, algorithm-assisted PI tool EmotionStream - that provides both contextual and emotional insights based on individual behaviors.
- An evaluation of the reliability of AER models and the role of context in stress prediction: we demonstrate that context, when augmented with emotional cues, shows promise for stress prediction using existing AER tools.

2 Related Work

We focus our literature review on CS student mental health, PI tools developed in the mental health space, and the underlying AER and Stress monitoring techniques used in the same.

2.1 CS Student Mental Health

As noted in the Introduction, the mental health of students in higher education is a growing concern. A study conducted among U.S. engineering students revealed that they are nearly ten times more likely to exhibit a high risk of serious mental health disorders compared to the general U.S. adult population [34]. A similar study reported that the prevalence of anxiety and depression symptoms was twice as high in CS students compared to other undergraduate students and 5-10x higher than in the general population [103]. Notably, students in Computing fields face the highest risk of mental health disorders among all engineering disciplines [34]. Multiple studies have explored the experiences of novice programmers in controlled and/or large scale classroom environments [15–18, 42, 59, 69, 115]. These studies revealed that students face both positive and negative experiences during programming [16–18]. Specifically, students experienced negative emotions like frustration more frequently during their first encounter with programming. Prior work has also pointed that negative academic experiences can affect student self-efficacy and academic outcomes [59, 69, 115]. Furthermore, imposter syndrome feelings-the doubt CS students had in their abilities after being burned out from academic tasks - were found to be prelevant in CS students [91]. Despite these efforts, the reasons why CS students experience elevated mental health issues are unknown. Recently, Ji et al. have investigated the reasons for stress among CS students, but the results are constrained only to students with pre-existing mental health conditions [52]. Our work explores the specific academic challenges that contribute to elevated stress and burnout in the broader CS student population.

In response to the growing mental health needs of university students, there has been a rapid proliferation of digital mental health tools. These efforts range from mobile sensing and digital phenotyping [72, 73] to the integration of Cognitive Behavioral Therapy (CBT) techniques and storytelling [45] to provide users with personalized and engaging approaches to address their mental health concerns. The declining mental health of the broader student population has also led to discussions about how to support student mental health within the CS Education community as recently as 2020 [2]. Research has explored the ways in which CS students may be able to reduce symptoms of anxiety and depression [37, 103]. A widely discussed study by Wang et al. through a continuous mobile sensing app studied the association between objective sensor data from smartphones and the mental well-being and academic performance of CS students [111]. More recently, Tran et al. examined whether gratitude journaling in an introductory CS course would reduce stress and improve life satisfaction in CS students but found no significant impact on stress levels between the control and intervention groups [107]. The Computer Science Education Community has largely explored the emotional responses to programming and specifically debugging tasks, however, our work is the first to understand stress and emotional responses from a mental health perspective outside large classroom environments in a student's naturalistic environment.

2.2 PI Tools for Mental Health

PI systems have been defined as multi-staged models that "help people collect personally relevant information for the purpose of self-reflection and gaining self-knowledge" [p. 2][67]. This model comprises five stages: 1) In the preparation stage, individuals identify the specific information to be recorded and establish the methods for data collection; 2) During the collection stage, data is systematically gathered; 3) In the integration stage, the collected data

is processed, combined, and transformed for analysis; 4) At the reflection stage, individuals engage with the data, exploring patterns and deriving insights; and 5) Finally, in the action stage, individuals decide how to apply the insights gained from the analysis to inform future actions or behaviors. PI systems, designed to gather and reflect on personal data to promote well-being and encourage positive behavioral changes, are commonly applied in managing mental health conditions such as mood [24, 77], emotions [48], and stress [1]. Visualization is commonly used as a medium for communicating insights in PI systems designed to facilitate reflection [4].

Early work on PI systems by McDuff et al. presented a multimodal system that continuously monitors users' valence, arousal, and engagement by utilizing various non-verbal and contextual signals [71]. In a recent study, Jorke et al. developed Pearl, a technology probe for machine-assisted reflection on personal data for worker well-being [53]. Another study by Kim et al. explored the design possibilities of incorporating prediction algorithms and explainability into PI systems to aid users in retrospective reflection using MindScope to examine how individuals interpret and use predictive algorithms for reflecting on stressors [57]. On the other hand, in a recent in-lab mixed methods study, Hollis. et al. examined how people understand and evaluate different algorithmic feedback about their personal emotional data [48].

Despite several tools being developed and tested in a variety of populations, including college students, there exists a research-to-practice gap in digital mental health; these tools developed and tested fail to achieve widespread adoption in real-world settings [66]. Specifically for college students, the key contributing factor to this gap has been suggested as the mismatch between tool design and their everyday experiences [50]. Specifically, they lack alignment of emotional states with the situation (activity type, application type, and privacy sensitivity of the activity) and temporal contexts, which was expressed as a recommendation by our participants to promote self-reflection, especially during academic tasks. Additionally, a recent study by Rooksby et al. revealed the human and ethical side of digital tracking and how it is critical to put student autonomy and self-determination at the heart of these approaches [90]. Inspired by their findings and similar others [56], we identify our population's preferences and privacy concerns for in-the-wild technological support to aid their mental well-being during academic work. Based on their openness and preferences, we designed EmotionStream - an algorithm-assisted PI tool that provides both contextual and emotional insights based on individual behaviors. Below, we detail the literature review on emotion recognition and real-time stress monitoring, which constitute integral components of EmotionStream.

2.2.1 Automated Emotion Recognition. Numerous methods have been explored to determine a user's emotional state, including assessing emotional states via physiological signals (e.g., heart rate, EEG, blood pressure) [29, 100], wearable sensor data (e.g., Microsoft Band or mobile phone) [88, 114], environmental data (e.g., lighting condition of the room, weather) [61], data directly reported by the user (via ESM or lifelogging) [39], and caretakers observations (e.g., parents in the case of infants) [36]. One such method is facial emotion recognition, which classifies user's emotional states based on facial cues from images. Facial Emotion Recognition has been widely used in both industry and research settings. Education and Learning Sciences have utilized it extensively in in-classroom and online and remote learning environments to promote learners' reflection of affect and also make teachers aware of the same. In the mental health domain, it has been used in self-tracking technologies, mood-based interventions, and various other applications. Further, Ruiz et al. showed that teachers and students can utilize early information about students' emotions to improve classroom results and learning outcomes [92]. In our work, We observed participants' emotions and context using EmotionStream - which utilizes two state-of-the-art facial emotion algorithms (output classes corresponding to seven basic emotions by Paul Ekman and Wallace Friesen [1]) and context logging. Participants also responded to ESM prompts to record their affect (as Positive, Negative, Neutral) in 20-minute intervals throughout each session. At the end of each session, a visualization dashboard showing temporal emotion cues along with context data is shown for promoting self-reflection.

However, existing facial emotion recognition systems are prone to various biases, such as those related to race, culture, and gender [113]. Therefore, deploying these systems in the real world may reinforce pre-existing biases. One reason for these biases is that the training sets do not represent the diverse population in the US (and this problem compounds when global representation is considered). In short, there is a difference between how the software codes for a particular emotion and what is going on in someone's mind - particularly for diverse groups [8]. The reliability of the most accessible emotion recognition frameworks is underresearched. Recently, Kaur et al. (2022) characterized the magnitude and type of misalignment between observed emotion and reported affect via a one-day study that combined AER tool predictions with diary entries [55]. In this line of research, as post-hoc secondary research analysis, we evaluate the accuracy of two state-of-the-art AER tools and investigate using these models for real-time stress prediction.

2.2.2 Real-time Stress Monitoring. Stress "occurs when demand exceeds the regulatory capacity of the organism" [30]. In a recent study, Ding et al. highlight that people often perceive their stress levels based on their own understanding of daily experiences [35]. While numerous technical methods for diagnosing stress help individuals interpret their stress levels, they often fall short of fully capturing the subjective nature of stress [49]. Early work by Adams et al. found that a self-report approach to detecting stress helps represent stress levels more accurately while also complementing algorithmic stress detection [1]. In a similar line of thought, Sanches et al. [94] argued that the key opportunity in designing stress management technologies lies in supporting individuals to reflect on their experiences to better interpret their stress levels rather than placing the focus on diagnosing stress.

A stressor is a social or emotional event that triggers a stressful response [28]. Stressors have been collected through various means, such as surveys [60], telephones [3], and smartphones [46]. Prior work has focused on understanding prevalent stressors and their role in depression [70], anxiety disorder [23], and PTSD [101], as well as in the broader population [3, 46, 60] for assisting users in managing their stress. For example, the DeepMood app [104] directed participants to input their moods and activities thrice a

[1] https://www.paulekman.com/facial-action-coding-system/

day to anticipate episodes of depression. As participants actively engaged in the mood-monitoring process through these apps, they observed an increase in emotional self-awareness.

Prior work has found that a large number of contextual features are linked to mental health conditions. At certain times of the day, including early morning [43] and nighttime [12, 105], anxiety and depression symptoms may become more pronounced. Additionally, Brown et al. [22] reported that stress in early adolescents is majorly impacted by their homework load. In our work, we periodically collect stressors and their associated stress levels to promote self-reflection and demonstrate the role of situational and temporal context in predicting momentary stress.

3 Needfinding Study

3.1 Method: Interviews with CS students

3.1.1 Data Collection. Our study participants were recruited using university list services and word-of-mouth. Our inclusion criteria were for a participant to be (i) 18 or older and (ii) a university student in CS. Prospective participants were invited to complete a preliminary survey. Once the research team received their response, the participants were sent a Calendly [2] link to provide their availability for scheduling an interview. Interviews were conducted from 27 December to 15 March 2023. Interviews lasted approximately 60 minutes (M=48) and were primarily conducted remotely over Zoom. Researchers and participants had their videos turned on during the interview, but only audio files were used for analysis. Participants who were uncomfortable turning on their video partook in an audio-only interview. We used Zoom's live transcription feature to automatically transcribe interviews and revised transcripts using Otter.ai [3], and manually verified them to improve accuracy. See our Supplemental Materials for more details on the interview protocol. Our participant demographics are as outlined in Table 1. Participants received $10 (USD) compensation for their time at the conclusion of the interview via an Amazon Gift Card or University Payroll.

3.1.2 Analysis. We used Braun and Clarke's thematic analysis framework to analyze interview data using a mix of inductive and deductive codes [20]. After conducting interviews with all 20 participants, the researcher independently coded the transcripts using open coding and identified emergent themes such as the role of academic coursework in Computing student mental health and the main challenges in managing their psychological well-being. The first and third authors coded a total of 20 transcripts. We used Dedoose [4] to code the interviews and achieved a Cohen's kappa value of 0.77 after two rounds of iteration (0.56 to 0.77). Throughout the analysis process, the team engaged in iterative and collaborative discussions to resolve disagreements and identify themes related to the challenges and needs of CS students. Our final codebook contained a set of 20 codes arranged into three high-level themes. See our Supplemental Materials for the final codebook.

[2]https://calendly.com/
[3]https://otter.ai/
[4]https://www.dedoose.com/

3.2 Findings: Challenges

3.2.1 Difficulties during Programming for Academic Work. Coding and debugging are integral components of the CS curriculum. Our participants all acknowledged attraction to the field due to their problem-solving inclination. Despite this, they (20/20) expressed debugging to be extremely stressful and frustrating. Participants identified a discrepancy between their expectations of successful code execution and the reality of encountering numerous errors, which they cited as the primary reason for facing difficulty with debugging (13/20). P13 noted:

> *Frustration is a big part of it [...] When I'm working on something, I kind of get stuck trying something over and over and over again. So I just keep getting frustrated. Naturally, you don't exactly know what you're trying to fix all the time [...] There were assignments in my sophomore year; I spent 14 hours debugging for a single class. These can be some pretty rough assignments. (P13)*

One-third of the participants noted a challenging contrast between grasping the logic behind coding concepts (e.g., loops) and the practical application of that knowledge demanded by assignments and labs, resulting in numerous errors in their code. Many reported that while the code may appear to work in theory or at certain stages, new bugs introduced during the debugging process often lead to additional frustration (9/20). Participants described a range of intense emotions during the process, including anger, frustration, restlessness, sadness, and panic. They often felt as though their efforts were in vain, leading to a sense of desperation and a desire to give up (9/20). P8, in her own words, describes her experience:

> *I've literally sat at my computer and cried. I remember I was taking [a Data Structures class] [...] I was just dead. I just started crying because I literally thought I was the dumbest person ever. I was like, I can't do this. And then I did that for 15 minutes. And then I settled down. [...] There's been times where I'm on a [time] crunch. And I literally have no time to do anything. So, I'm frantically going through things. And I'm just not even feeling anything. I'm just reading and typing. And I'm like, what isn't working? And I'm reading a million times because I have to get it in in a few hours. And then you're happy when it finally works, and you're really excited. So I've just been through every emotion while doing that. (P8)*

Other challenges raised by participants during coding include difficulties in remembering syntax (P4, P8), confusion in deciphering library documentation (P1, P8), and navigating multiple programming languages and IDEs (P4, P8). Several (6/20) noted that coding courses were markedly different from other non-CS courses they may have taken, primarily due to the elevated stress levels associated with debugging, their time-consuming nature, and the imperative to complete several of them within tight deadlines.

3.2.2 Imposter syndrome. Imposter syndrome (misrepresentation of self in academic life as defined by Bothello et al. [19]) was reportedly a common issue in participants. Over half (12/20) self-reported

ID (Age, Gender)	Ethnicity	First-Gen	Student Status	Attended Counseling?	Time spent on Computer (hrs)	Tool study participant?
P0 (21, Male)	White	No	Domestic, Graduate	No	8	Yes
P1 (25, Male)	White	No	Domestic, Graduate	Yes	12	No
P2 (30, Female)	Asian	No	International, Graduate	No	3	Yes
P3 (22, Male)	Asian	No	International, Graduate	No	5	Yes
P4 (21, Female)	Asian,White	No	Domestic, Undergraduate	Yes	12	No
P5 (24, Male)	Asian	No	International, Graduate	Yes	12	No
P6 (29, Male)	Asian	No	International, Graduate	Yes	12	No
P7 (23, Non-binary)	White	No	Domestic, Undergraduate	Yes	10	No
P8 (22, Female)	White	No	Domestic, Undergraduate	Yes	5	No
P9 (26, Female)	Asian	No	International, Graduate	Yes	8	No
P10 (19, Female)	Asian	Yes	Domestic, Undergraduate	No	5	No
P11 (21, Male)	Asian	No	Domestic, Undergraduate	No	1-5	Yes
P12 (25, Male)	Asian	Yes	International, Graduate	No	8	Yes
P13 (21, Male)	White	No	Domestic, Undergraduate	Yes	4.5	Yes
P14 (20, Male)	White	No	Domestic, Undergraduate	No	Not reported	No
P15 (22, Female)	Asian	Yes	Domestic, Undergraduate	Yes	8	Yes
P16 (20, Male)	African American	Yes	Domestic, Undergraduate	Yes	6	No
P17 (18, Female)	Asian	No	Domestic, Undergraduate	No	5	No
P18 (21, Male)	White	No	Domestic, Undergraduate	No	4	Yes
P19 (18, Male)	White	No	Domestic, Undergraduate	Yes	4	Yes
P20 (20, Female)	White	No	Domestic, Undergraduate	N/A	6	Yes
P21 (25, Female)	Asian	No	International, Graduate	N/A	8	Yes
P22 (27, Female)	Asian	Yes	International, Graduate	N/A	2	Yes

Table 1: Participant demographics

experiencing imposter syndrome, aggravated by the constant pressure to prove themselves in the CS field. While participants acknowledged that this was common across disciplines, they perceived imposter syndrome as more pervasive in CS due to the rapid pace of advancements in and high expectations of the field. Participants mentioned associating negativity with themselves, such as being labeled as a "bad programmer" (P15) or doubting their aptitude in CS (P4). P14 also highlighted how imposter syndrome created a competitive environment when he said,

> *Yeah, the biggest challenge I've seen is the pressure to keep proving yourself in this field; it's very competitive, especially given that it is continuously growing. And being able to set yourself apart from your peers and let an employer know, hey, I'm better than the guy next to me. I feel that the pressure of trying to beat the person in front of you is the biggest challenge many CS students face today. (P14)*

3.2.3 Lack of Self-Awareness Mental States. Participants experienced intense emotions and stress during their academic work, especially debugging, as noted in Section 3.2.1. However, one-third (7/20) of participants reported that prolonged efforts to fix their code led to a lack of self-awareness of their stress and emotions, resulting in a cycle of sleep deprivation and burnout. In the absence of stress-monitoring practices, participants sometimes relied on their own judgment to identify moments of stress. For instance, P10 shared her practice of taking breaks every hour during coding and debugging sessions, whereas P7 mentioned her intention to set a timer for 30 minutes and do a self-check about how she is feeling but was not successful in implementing it. Although participants (14/20) acknowledged the importance of breaks for their mental well-being, they struggled to adhere to them.

Frequent lack of recognition of stress often led to feelings of burnout. Most participants (16/20) reported experiencing burnout during their CS program. Several (5/20) reported experiencing burnout during exam times at the end of semesters when their stress levels peaked. While most were certain about their experience, some were unsure until their therapists helped them recognize their burnout (3/20). Notably, one mentioned requiring intervention from their professor to take a break (P3). P3 summarizes her experience as:

> *Sometimes I think I don't understand that I'm stressed. For example, last semester, I had an Android project to do. So. I finished most of the coding part. But at that time, I didn't realize that I was not taking breaks that much because I love coding. [...] But when it was exam time, it was so difficult for me because, for example, I was seeing 'not' instead of 'hot,' but I didn't understand that maybe I was stressed.* (P03)

Participants noted that pre-existing challenges from other sources, including health and personal life, exacerbated their burnout experience and vice versa. Participants felt that extreme burnout was difficult to recover from. For example, P7 started to feel as though her burnout was causing her to lose her memory.

3.2.4 Other challenges. Apart from the detailed challenges in the above sections, participants mentioned suffering from academic procrastination, leading to increased stress levels, especially during tight timelines. Additionally, they had existing mental and personal challenges that, when combined with academic challenges, made them want to leave the program altogether.

3.3 Findings: Technological Support Preferences

In the sections to follow, we first explore whether participants are open to using technology to self-manage their psychological well-being. Next, we detail the informational elements and visualization features participants envisioned to aid their mental well-being while performing academic work.

3.3.1 Adoption of Technology. As shown in Table 1, more than half of the participants (11/20) in our study sought professional help to manage their mental health. Regardless of seeking or being in professional care, participants emphasized the importance of self-resilience in managing their symptoms (P1, P14, P15). For example, P1 said: *"I have visited counselors before, but at the end of the day, whatever I'm going to do is going to be the best thing to help myself."* Although participants emphasized the importance of resilience, none reported using specific tools to support their mental health. Instead, they relied on general productivity tools, such as digital note-taking and to-do list apps, rather than dedicated mental health support systems. When asked about their willingness to adopt a computer-based PI tool, most participants indicated that they were open to using such a tool to facilitate self-reflection and manage their psychological well-being during academic work (16/20). P2 mentioned: *"Not everyone comes with the same background; very few people have [mental health] experience and can manage better, but many cannot, so a tool will certainly help."*(P2). Notedly, P18 mentioned:

> *...[a tool] might not be immediately useful, but you can probably recognize some patterns out of it. And maybe use that to diagnose the [mental health] issue. Maybe if you went to therapy, it'd be an interesting tool to show and the [therapist] can kind of pinpoint a better solution to your problems, as opposed to without a tool. (P18)*

Participants extensively reported using their computers for most of their academic work over smartphones or tablets and expressed a strong preference to be monitored via their computers.

3.3.2 Real-time stress and emotional state monitoring. Most participants (18/20) expressed a preference for monitoring their mental states during academic tasks. From the interviews, participants expressed a preference for two types of monitoring: 1) Emotional state monitoring (13/20) and 2) Stress monitoring (14/20). For instance, P7 highlighted the value of emotional state monitoring, stating:

> *I mean, for at least people who are less conscious of their emotions, it makes sense to use facial emotion recognition. I know it's not like the most accurate technology, but even a little bit of info into how they seem to be feeling while they're working on stuff could definitely be an eye-opener for some people.* (P7)

Likewise, for stress monitoring, P17 noted:

> *I feel like they can really benefit from some technology that helps them monitor their stress levels in some*

> *way; that can really help them be self-aware of how much stress they're having and take care of themselves. I would explore it [such a tool]* (P17)

All participants identified two potential benefits of monitoring stress and emotional states: *1) Promote awareness of their feelings during computer-based academic tasks and 2) Enable them to self-manage (e.g., taking breaks when feeling consistently frustrated).* Participants (18/20) emphasized stress and mood-based breaks, and P18 reflected this when they said:

> *I think one of the big things is, if I would notice that I'm at a volatile state where it's not too good, I guess it might be an indication that it is time to take a five-minute break, [...] because I feel like with programming, you can kind of get stuck in this idea that you have to get it done all in one go. And that's usually just not the case. (P18)*

3.3.3 Contextual monitoring. In addition to monitoring their emotional states and stress, all participants expressed a need to monitor the contexts in which their emotional states and stress levels deviated from what was normal to them. Participants (7/20) identified the key contextual factors associated with their stress levels and emotional states to be: *1) Task-level data, 2) Keystroke metrics, and 3) Other contextual factors such as time of day*

Task-level data: Participants expressed the need to identify the sources of their stress and emotional responses, particularly the activities that triggered these reactions. They also noted that a detailed breakdown of the time spent on each activity and the applications used would be beneficial in determining when to take breaks.

Keystroke metrics: Participants, in addition to measuring task-level metrics, showed an affinity to track any behavioral trends revealed by keystroke data, such as typing cadence. For example, P12 states:

> *So I would say, this was something that came to my mind based on the way you were typing the keys sometimes. Because I would tap them hard or tap them fast if I'm stressed or doing some programming tasks, and I'm not able to achieve that task, or maybe bang that space bar, identify those patterns, if possible.* (P12)

Other contextual factors: In addition to the previously mentioned metrics, four participants recognized the importance of including the time of day they were working, noting that late-night coding sessions were associated with increased stress levels.

3.3.4 A visual dashboard summarizing emotional trend. When asked how they would like to make sense of the data collected from these logging components listed above, participants (15/20) came up with various ideas, all of which had to be a summary of how they felt during the task and for how long.

> *Maybe some kind of symbol representing how you're feeling, like a little smiley face or a frowny face or, you know, stressed out a face or something that would maybe alert you if you were getting too stressed out or something and say, go take a 10-minute break. But not too in your face. Because if you can't take a break, you know, you can't take a break, you'd have to be able to easily shut it off or something. But I think something like that would be would be very beneficial.* (P1)

Participants expressed a desire for a dashboard that includes a timeline of their emotional states along with a summary of task-level data each time they engaged in academic work.

3.3.5 Preserve their privacy. Participants (16/20) preferred passive tracking, avoiding identifiable data such as facial cues, audio, and their private application activity to be recorded. Participants emphasized the importance of storing and processing collected data locally, avoiding the need to send data to the cloud. They also highlighted the need for transparency regarding what data will be collected, where it will be stored, and how it will be processed. When sharing data, participants preferred it to be done anonymously and in aggregate form, ensuring that no individual could be identified. Participants further emphasized that the technological solution's primary purpose should be to help them manage their stress and emotional responses, with enrollment remaining voluntary rather than mandated by the university. Four participants expressed concerns about the potential negative impact of monitoring, describing it as a "double-edged sword" that could increase their stress due to the feeling of being constantly observed.

3.4 Design Goals

We distill our interview study findings into the following Design Goals:

- G1: Integrate computer-based mental health tools into academic environments
- G2: Continuously track emotional states
- G3: Track stress levels and associated contexts
- G4: A visual dashboard summarizing emotional responses along with context data
- G5: Maintain data locality and privacy

4 EmotionStream

As described in the previous section, our formative need-finding work helped us determine a set of design goals that act as a foundation for building a tool for CS student mental well-being. With these goals in mind, we designed and built EmotionStream—a computer-mediated algorithm-assisted PI tool. Next, we present a user scenario to describe how EmotionStream can be used, followed by features and the system's implementation.

4.1 User Scenario

Alexa is a university CS student at University X who chose the major due to her love for coding and problem-solving. She recently started taking core CS courses and quickly began to realize that she is doing multiple coding assignments on a weekly basis. She started an assignment on a Saturday morning and went on till the night, although she thought she would break for lunch and dinner. Her code that night had more errors than she had begun with. Two weeks later, she realized how burned out she was trying to submit 3 such assignments on time. She realizes she lost sleep and was under constant stress over these assignments.

Looking to reduce her stress and understand her emotional responses to academic tasks, she starts using EmotionStream. She

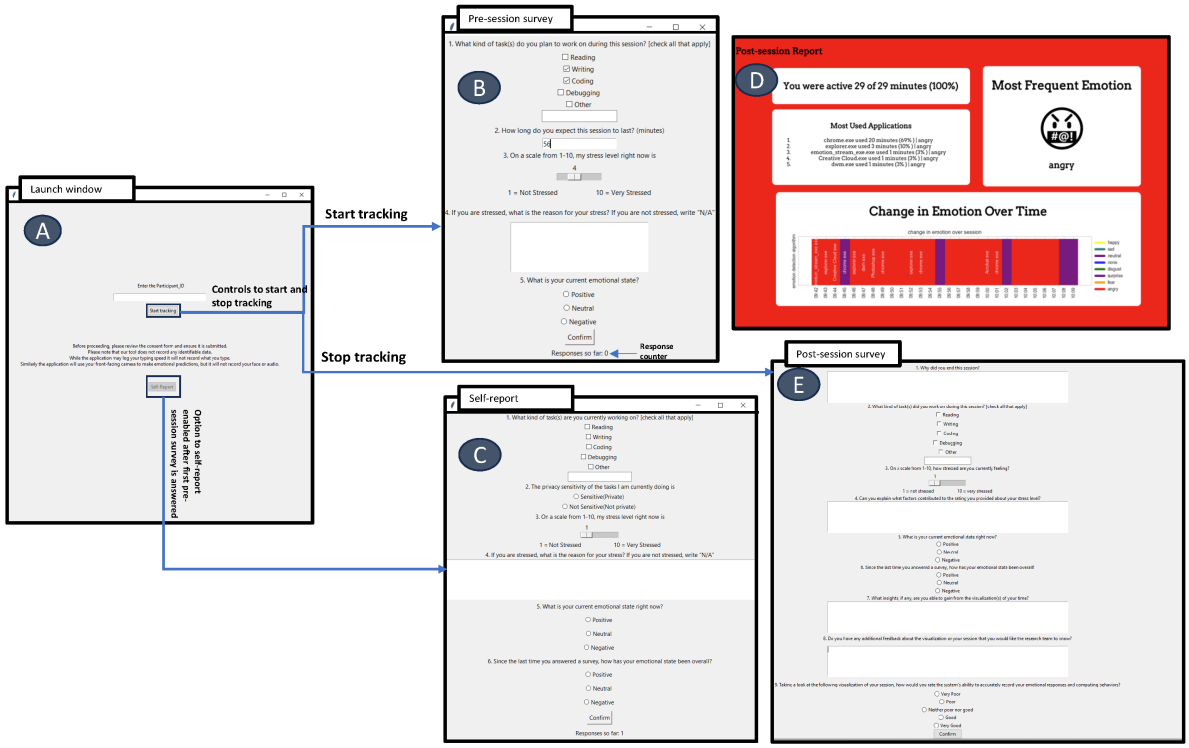


Figure 1: Interface Design of EmotionStream. (A) The Consent and Initialization View, (B) Pre-Session View, (C) Contextual Self-Reflection View, (E) Post-Session Insight View, along with visualization dashboard (D)

notices that she has the option to start tracking her data once she launches the app and clicks 'Start Tracking'. Once she clicks this button, she is prompted to respond about what activities she will be working on, how long she is planning to work on, her current affect, stress, and the reason for that stress level. She begins working on her assignments and observes that her camera has been enabled. After 20 minutes, she is prompted to reflect on how she is feeling, how stressed she is, the reason for her stress, and what activity she was doing. After her first prompt, she notices these prompts every 20 minutes and answers them. When it is time for her to go to class, she hits 'Stop Tracking' on the application and finds her camera turn off. She is shown a visualization dashboard with her most prominent emotional state during that session, her top 5 applications and emotional states associated with them, how active she was during that session, and the temporal alignment of her emotional state during that session. When she views this visualization, she responds to a few additional questions, this time regarding her feedback about the system and the dashboard.

One day later, she logs into EmotionStream to find that this is her second time using it, and she has responded to 10 surveys so far. She browses the folder in which her data is saved to find all the data from her previous session and the dashboard as well. In this folder, she finds all her logs related to temporal emotional states, peripheral data (keys per minute, clicks per minute, key press duration), her prompt responses, her feedback about the system, and the visualization for each session.

4.2 EmotionStream Interface Design

In this section, we outline the interface of EmotionStream and the different views. The Consent and Initialization View prompts users to review the consent form and provides controls to initiate or halt tracking. The Pre-Session View facilitates session preparation by enabling users to log their planned activity, intended duration, and current stressors and emotions. During the session, the Contextual Self-Reflection View is triggered every 20 minutes to encourage self-reflection on activities, stress, and emotions, while also allowing users to self-report additional ESM responses as needed. The Post-Session Insight View presents a survey alongside the system-generated visualization dashboard. This view supports self-reflection and taking free-form notes to capture their observations and lessons learned. The EmotionStream interface, depicted in Figure 1, was developed using Python's TkInter module.

4.3 Features of EmotionStream

We built EmotionStream with the aim of addressing design goals (G1-G5) presented in Section 3. We describe the three primary components in the following: passive emotion recognition and context logging, ESM surveys, and the front-end visualization dashboard design.

4.3.1 Automated Emotion Recognition (G1, G2, G5). Findings from Section 3.3.2 revealed that participants wanted to continuously monitor their emotional states (G2) on their computers when working on academic tasks (G1). Analyzing facial expressions is a widely used method to continuously detect emotional states in computer vision. To do this, we employed two state-of-the-art open-source facial emotion recognition models, DeepFace [97–99] and Residual Masking Network (RMN) [85]. Each of these models uses its own facial recognition packages. DeepFace uses a hybrid facial recognition model by wrapping multiple models like VGG-Face, Google FaceNet, OpenFace, and Facebook DeepFace [97–99]. RMN uses an OpenCV standard face detector model [85]. Every model listed captures facial features at the frame level. Our application records the timestamp at which the frame was identified, the emotion detected, the probability of the detected emotion, and the probability list of the seven emotion classes (angry, disgust, fear, happy, sad, surprise, neutral) from both models. To maintain privacy as per participant preferences (G5, Section 3.3.5), no personally identifiable data, such as the actual face or facial cues, were recorded.

4.3.2 Context Logging. Our formative study revealed participants wished to capture the context in which they are at heightened moments of stress (Section 3.3.3). Therefore, we defined context as peripheral logging(keystroke, mouse), activity type, and application usage.

Key Logger (G3, G5). The keylogger module takes real-time updates from the user's keyboard interactions using the "keyboard" Python module [5]. Our application records the timestamp of the action, the type of action (UP, DOWN), and the type of key (KEY, SPACE, BACKSPACE) rather than the actual key pressed to maintain privacy (G5, Section 3.3.5). We use the logged data to compute attributes such as keystrokes per minute, average keypress length, average delay between key presses, and most keys pressed in a given interval.

Mouse Logger (G3). The mouse logger takes real-time updates from the user's mouse interactions using the PYNPUT library [6]. Our application records the timestamp of the action and the type of action (MOVE, CLICK, SCROLL). Move and click actions write the coordinates of the mouse, while scroll actions log the scroll vector, indicating the velocity of the scroll. We use the data from the mouse logger to compute attributes such as clicks per minute, total mouse movement, average mouse speed, average scroll velocity, changes in scroll direction, mouse click length, and mouse click delay.

Application Logger (G3, G5). To further capture the context of participant's academic work (G3), the Application Logger captures their interactions with applications. This component takes updates on one-minute intervals that log the timestamp, the application in the foreground, and the number of applications in the background. We tracked the top five applications most frequently in focus during each user session at one-minute intervals by analyzing CPU usage at specific timestamps. We use these data to calculate the time spent on each application. We do not record the screen of the participant or collect any data that is personally identifiable (G5, Section 3.3.5).

4.3.3 Experience Sampling Surveys (G3). EmotionStream prompted users with a Pre-session View at the beginning of each session, a Contextual Self-Reflection View at 20-minute intervals during their session, and a Post-session View at the end of each session. Q1, Q2, Q4-Q6 from Table 2 are displayed to the user in the Pre-Session View as soon as they launch the application and start tracking. The ESM surveys during the session in the Contextual Self-Reflection View consist of Q1, Q3-Q7 from Table 2. The ESM prompts during the session were triggered under one of three conditions: (i) a strong (with >90% probability) negative emotion predicted by the AER models, (ii) a random interval N (where N is an integer between 1 and 20 minutes) within the 20-minute block, or (iii) a manual self-report initiated by the participant. These parameters were refined through a pilot study with five participants prior to deployment to minimize response burden.

4.3.4 Personal Informatics Dashboard (G4). The Post-session Insight View featured a dashboard visualization summarizing session data, including the proportion of time spent actively engaging with each application, the dominant emotion associated with each application, the most prevalent emotion throughout the session, temporally aligned emotional states, and peripheral activity. A user was deemed active during a one-minute interval if they interacted with their peripherals (e.g., typing or using the mouse); inactivity was recorded in the absence of such interaction. The emotion linked to each application was identified by selecting the most frequent emotion during its use at the minute level. A sample of the dashboard is presented in Figure 1.

5 Evaluation of EmotionStream

We structure the evaluation of EmotionStream into three subsections: 1) Self-Reported Measures and Contextual Associations: We summarize self-reported affect, stress, and stressors, highlighting statistically significant associations between stress, self-reported affect, and user context (mouse and keystroke interactions, time, application use, and activities). 2) Tool Acceptability: We evaluate the acceptability of EmotionStream through two metrics: tool engagement and rating responses from a post-session survey. 3) Post-Hoc Evaluation: Lastly, we evaluate the accuracy of AER models and predict stress utilizing self-reported affect and context.

5.1 Methods

We reached out to the same participants we interviewed during our needfinding study and also conducted an additional round of recruitment through university list services and word of mouth. 10 participants were previously involved in our earlier study, and we gained 2 new participants. We distributed an intake form via Qualtrics, and based on the interest expressed, we sent them installation instructions. One additional inclusion criterion was for our participants to use a Windows OS-based personal computer (laptop

[5] https://pypi.org/project/keyboard/
[6] https://pypi.org/project/pynput/

I. Context	Data Type
Q1. What tasks are you currently working on?	Categorical [Reading, Coding, Debugging, Writing, Other]
Q2. How long do you expect your session to last?	Numerical
Q3. What is the privacy sensitivity of the task you are currently working on?	Categorical [Private, Not private]
II. Stress	**Data Type**
Q4. My stress level right now is	Likert (Scale: 1-10)
Q5. Reason for stress is	Free-form Text
III. Emotion	**Data Type**
Q6. Current emotional state right now is	Categorical [Positive, Negative, Neutral]
Q7. Emotion since the last time you answered the survey	Categorical [Positive, Negative, Neutral]
IV. Feedback	**Data Type**
Q8. What insights were you able to gain from the visualization?	Free-form Text
Q9. Do you have any feedback about the tool or visualization for the research team?	Free-form Text
Q10. How would you rate the system's ability to accurately capture your emotions and computing behavior?	Likert (Ordinal)

Table 2: Questions from pre-, in-situ and post-surveys

or desktop with a webcam) because we had created a Windows-only application. At the beginning of the study, we gathered information from our participants regarding demographics (e.g., age, gender, student status). In particular, we also collected the average time they spend using their computers for academic coursework daily. To collect data about the role of academic coursework in CS students' well-being, asked participants to install EmotionStream. During the study, each participant was asked to use our tool for at least one hour per day for a week while performing academic tasks.

5.2 Analysis

The AER models used in this study predict emotions as one of seven classes: fear, disgust, anger, happy, sad, surprise, and neutral. Following conventions in Affective Computing research [86], we categorized these predictions further into: positive (happy, surprised), negative (angry, disgusted, fearful, sad), and neutral. From here on, we use the term affect for self-reported emotion ground truth. To determine the alignment of AER model predictions with self-reported affect from ESM surveys, we aggregated dominant emotions from frames over one-minute intervals [32, 54]. We calculated the percentage dominance of each emotion across the study duration, generating individual emotion profiles for participants, as defined by [55]. For transparency, the distribution of all seven emotion classes from both AER models is depicted in Figure 4. However, only Positive, Negative, and Neutral were included in the primary analysis for consistency with self-reported data. We extrapolated minute-level data by extending the last reported affect across the interval between consecutive ESM responses.

5.3 Findings

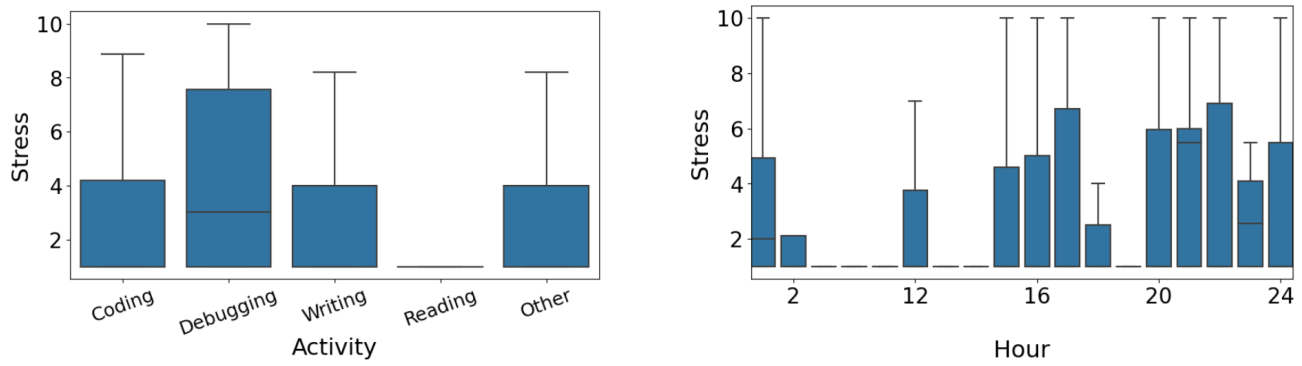
Over a 5-day data collection period, 12 participants generated 408 survey responses. Each participant averaged 34 responses (Median=25, SD=24.6). They completed an average of 11 sessions (Median=9, SD=4, Range=5-18) and an average of 11.6 hours logged (Range=5-27). Survey responses per participant ranged from 7 to 95. They reported engaging in 20 distinct activities, with the most frequent activities being coding, debugging, reading, and writing. The average session length was 1.13 hours. We excluded timed-out prompts that expired after a 5-minute duration.

5.3.1 *Self-reported Affect and Stress.*

Self-reported Affect. Participants' affect were collected through the Experience Sampling Method as described in Section 4.3.3. During the study period, 73% of the self-reports were Neutral, 20% Negative, and 7% Positive. Among the participants, 42% experienced two out of three affect. Specifically, 17% experienced only positive and neutral, while 25% experienced only negative and neutral emotions. When comparing male and female responses, females reported a greater proportion of positive emotions (23%) as opposed to males (2%). Due to the lack of self-identification of other genders, we only report Male and Female comparisons.

Self-reported Stress. The study collected stress level data from self-reports at 20-minute intervals. Participants' average stress levels across sessions ranged from 1.7 to 5.8. When stress levels were observed across activities, reading was associated with overall lower stress values, and debugging was associated with the highest stress across participants. Figure 2a displays the median stress levels of different activities, revealing that certain activities, such as debugging, have more than 25% of the ESM responses with stress values above 7.5. Coding, on the other hand, was associated with more than 25% of its ESM responses above 4, and the maximum stress value reported is 10.

Self-reported Stressors. Through the tool, at every 20-minute interval, participants reported various stressors during the study, ranging from personal health reasons to exams. Here, we provide an overview of the stressors for two reasons - first, to provide verification that participants raised challenges consistent with needfinding



(a) Activity wise stress data

(b) Hour of day wise stress data

Figure 2: Activity and hour of day stress levels

interviews and to provide context for the additional analysis that follows. Academic stressors emerged from 1) thinking about potential things to do in the future to fulfill their academic requirements, such as a deadline to finish an assignment, having multiple assignments to complete within X time, and having upcoming exams and project submissions, finding an internship to stand out 2) ongoing project work, assignments, timed quizzes, debugging code, unable to make progress 3) worrying about performance in a past exam, quiz, or coursework in general. Participants also reported additional stressors from personal issues such as health (eg. headaches, eye stress), and personal commitments taking up their time

Association of Affect and Context with Self-reported Stress. To better understand stress, we considered stress and its correlation with affect and context features. In particular, we used Spearman's correlation coefficient, reported correlation coefficients, and p-values for each. There exists a negative and significant correlation between self-reported affect and stress (r=-0.36, $p < 0.001$). Additionally, a negative correlation was observed between the number of applications and stress levels (r=-0.28, $p < 0.001$). The hour of the day the student was working on their academic coursework was positively and significantly correlated with stress (r=0.2, $p < 0.001$). Similarly, the correlation of stress to mouse click speed (r=0.06, $p < 0.001$), keystroke speed (r=-0.1, $p < 0.001$), and key press duration (r=-0.039, $p < 0.001$) were weak but significant. We did not find significance in the number of unique keys pressed. Participants reported 4 major activities in their academic coursework. We report correlations of stress with the contextual features and affect during these activities.

a. Coding. Stress was negatively and significantly correlated with self-reported affect (r=-0.74, $p < 0.001$) and with the number of open applications (r=-0.38, $p < 0.001$), key press speed (r=-0.11,$p < 0.001$), key press duration (r=-0.15,$p < 0.001$), and unique keys pressed (r=-0.1,$p < 0.001$). Stress was positively correlated with the hour of the day (r=0.17, $p < 0.001$) and mouse click speed (r=0.1, $p < 0.001$).

b. Debugging. Stress was negatively and significantly correlated with self-reported affect (r=-0.34, $p < 0.001$), number of open applications (r=-0.18, $p < 0.001$), and key press speed (r=-0.1, $p < 0.001$). A positive correlation was found for mouse click speed (r=0.1, $p < 0.001$), and no significant correlation was found during debugging for the key pressed duration and unique keys pressed.

c. Reading. Stress was negatively correlated with the number of open total applications (r=0.17, $p < 0.001$), key press speed (r=-0.06, $p < 0.001$), unique keys pressed, and keys pressed duration had a weak correlation. No significant correlation was found between mouse click speed and self-reported affect.

d. Writing. Stress was negatively and significantly correlated with the total number of open applications (r=-0.4, $p < 0.001$) and self-reported affect (r=-0.44, $p < 0.001$). It was positively correlated with mouse click speed (r=0.15, $p < 0.001$). There was no significant correlation with keys pressed speed and a weak correlation with other keyboard features.

5.3.2 *Acceptability.*

Tool rating. We analyze the responses to 3 feedback questions about the tool during post-session surveys. Figure 3 shows the percentage of system ratings.

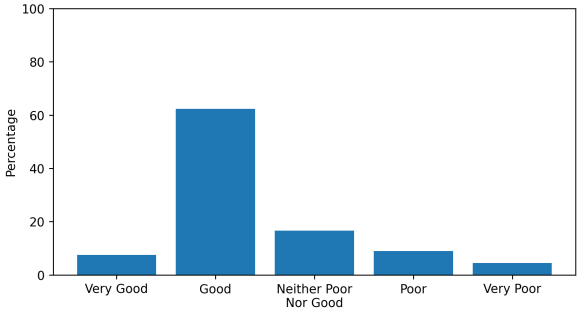
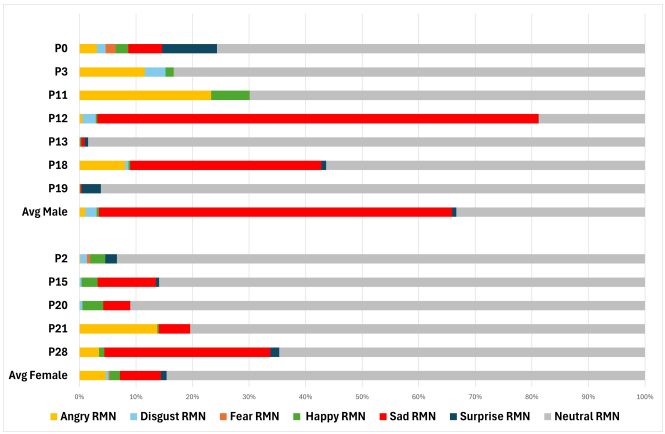
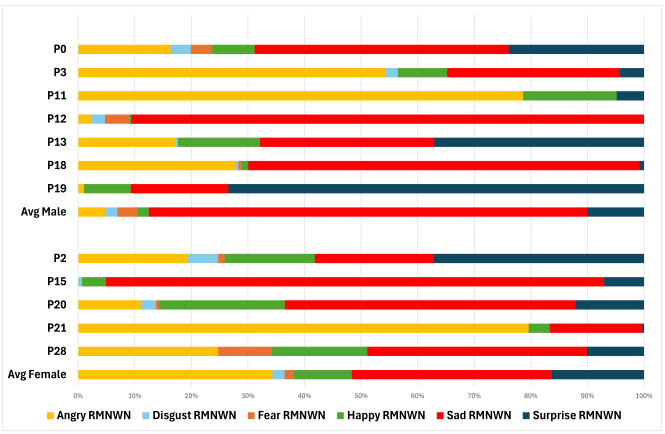


Figure 3: System Ratings

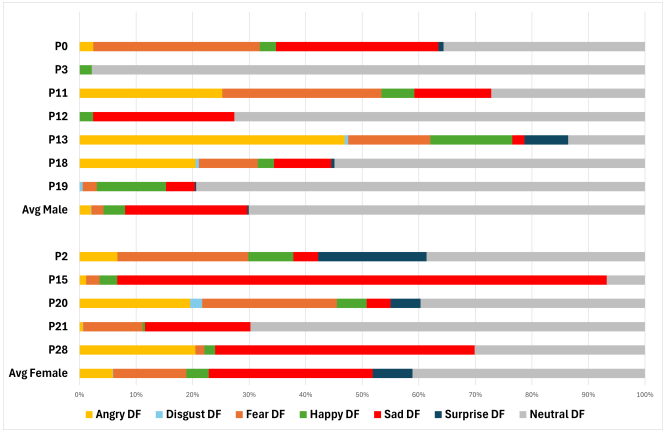
The scores indicate that participants were very satisfied with the app as EmotionStream received a rating of "Good" and "Very Good" from 60% and 10% of the survey responses, respectively. Open-ended feedback from participants explains their sentiments.



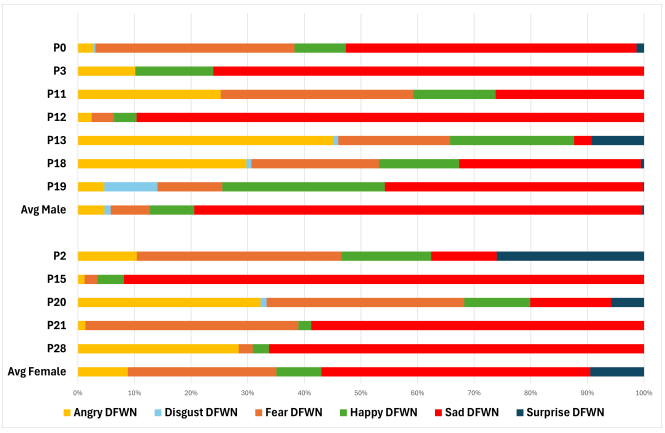
(a) Dominant emotion profile of RMN model aggregated at 1-minute intervals



(b) Dominant emotion profile after removing Neutral label of RMN model aggregated at 1-minute intervals



(c) Dominant Emotion profile of DeepFace aggregated at 1-minute intervals



(d) Dominant emotion profile after removing Neutral label of DeepFace aggregated at 1-minute intervals

Figure 4: Emotion profiles of all participants

Participants found it helpful to reflect on their affect and stress at frequent intervals. For example, P18 mentioned, *"I think the main reason for the reduction in my stress was because often I was asked about the nature of how I was feeling and why. Grappling with that question led to progress towards a better emotional state"*. Participants found the visualization at the end of their sessions to be helpful. For example, P13 mentioned *"I did not realize how much I switch between different programs, which is easily seen with the visualization."* However, some participants also provided feedback that they would prefer to see more explanation for the data shown on the dashboard. For example, *"It would be useful to see the amount of activity associated with each application."*

Tool Engagement. We assessed tool engagement through two key metrics: 1) the total hours participants interacted with the tool and 2) the number of surveys (ESM responses) completed by participants. Although participants were only required to use the tool for 5 hours, the average tool usage time was 12 hours, more than double the required time. The highest engagement time was 29 hours. For Experience Sampling Method (ESM) responses, the response rate to the prompts was 92%. Three participants (P0, P11, and P13) had a 100% response rate, while the lowest response rate was 77.3% for P28. Additionally, two participants voluntarily chose to self-report their emotional states during the study, with P11 self-reporting in 39% of the sessions, and P18 in 16%. These findings suggest high participant engagement, both in terms of required use and voluntary interaction, indicating that the tool was well-received and consistently utilized for self-monitoring.

5.3.3 Post-hoc evaluation. Lastly, we evaluate the accuracy of AER models and predict stress using self-reported affect and context.

Accuracy of AER models: We compared the two AER model predictions aggregated at the minute-level with self-reported affect. For comparison, only instances with complete data from all three sources (i.e., two model predictions and ESM survey responses) were included.

1. Residual Masking Network. The RMN model's overall accuracy for

RMN Predicted Values

Actual	Positive	Negative	Neutral	Total
Positive	**31**	188	1250	1469
Negative	69	**2170**	2121	4360
Neutral	176	6913	**5631**	12720
Total	276	9271	9002	18549

DeepFace Predicted Values

Actual	Positive	Negative	Neutral	Total
Positive	**83**	413	973	1469
Negative	250	**1172**	2938	4360
Neutral	787	4459	**7474**	12720
Total	1120	6044	11385	18549

Table 3: Classification values for predicted vs. self-reported affect. Left: Contingency matrix for affect predicted by RMN. Right: Contingency matrix for affect predicted by DeepFace

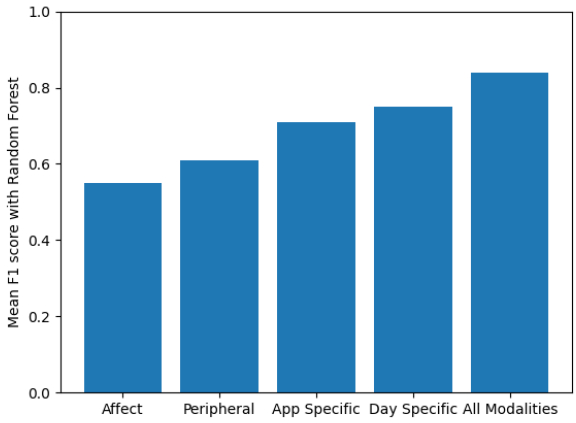
classifying Positive, Negative, and Neutral affect is 42%, with precision at 0.49, recall at 0.42, and F1-Score at 0.43 (derived from Table 3). Additionally, a chi-square test between the model's prediction and self-reported affect revealed a significant difference between the two ($\chi^2(2, N=18549) = 912.17$, $p < 0.001$, $V = 0.15$). A majority of Positive and Negative affect were classified as Neutral, resulting in the low accuracy of these two classes (2% and 49%, respectively). A comparative analysis of Males and Females revealed accuracies of 37% (Precision: 0.53, Recall: 0.37, F1-Score: 0.39) and 55% (Precision: 0.45, Recall: 0.55, F1-Score: 0.47), respectively.

2. DeepFace: On the other hand, the DeepFace model's overall accuracy for classifying Positive, Negative, and Neutral affect is 47%, with precision at 0.50, recall at 0.47, and F1-score at 0.48. Further, a chi-square test between the model's prediction and self-reported affect revealed a lower chi-square value in the two distributions ($\chi^2(2, N=18549) = 123.37$, $p < 0.001$, $V = 0.05$) when compared to that of RMN, signaling a lower difference in the two values. The accuracy of Positive and Negative affect is 5% and 27%, respectively. A comparative analysis of Males and Females revealed accuracies of 54% (Precision: 0.55, Recall: 0.54, F1-Score: 0.54) and 31% (Precision: 0.35, Recall: 0.31, F1-Score: 0.30), respectively.

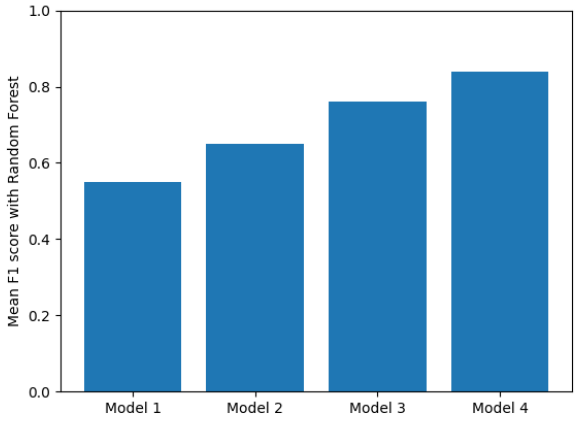
3. Comparing RMN and DeepFace: In a chi-squared test-based comparative analysis, a significant difference was found between the predictions of each model ($\chi^2(2, N=18549) = 745.41$, $p < 0.001$, $V = 0.14$). Further, we can see that the RMN model classifies Females as expressing more positive affect, while DeepFace predicts Females as expressing more negative affect compared to Males, as seen in Figure 4.

Predict stress from affect and contextual cues: Three different analyses were conducted to predict stress. For all the analysis, we trained independent Random Forest classifiers for each user and reported the average F1 scores across individuals. The modalities (features) used in the classifiers are self-reported affect, peripheral (mouse click speed, key press speed, key press duration, and unique keys pressed), app-related (foreground application, number of background applications, and activity type), and day specific (hour and time of day, day of week).

1. Stress prediction from individual modalities. Figure 5a shows the F1 scores when Random Forest Classifiers were trained with each individual modality to predict stress. F1 scores from each modality were 0.55 from self-reported affect, followed by peripherals (0.61), app-specific data (0.71), and day-specific data (0.75). The best F1 score was 0.84 when all the modalities were used to predict stress.



(a) F1 scores from individual modalities



(b) F1 scores from forward stepwise feature selection

Figure 5: Average F1 scores. (Left) Peripheral data: key press speed + unique keys pressed + key press duration + scroll velocity + scroll action; Application: activity type + current application + the number of applications open; Day specific features: hour of day + time of day + day of week. (Right) Model 1: affect; Model 2: affect + peripheral data; Model 3: Model 2 + activity data; Model 4: Model 3 + day specific features

Our secondary analysis using RMN and DeepFace affect as a modality to predict stress yielded F1 scores of 0.50 and 0.50, respectively.

2. Forward stepwise feature selection. Figure 5b shows the F1 scores when forward stepwise feature selection was performed to predict stress. Model 1, comprising only self-reported affect, yielded an F1-Score of 0.55. Model 2, an incremental addition to Model 1 with peripheral data, yielded an F1 score of 0.65. Model 3, an incremental addition to Model 2 with app-related modality, yielded an F1 score of 0.76. Model 4, an incremental addition to Model 3 with day-specific modality, yielded an F1 score of 0.84. Secondary analysis comprising of RMN and DeepFace predicted affect instead of self-reported affect in Model 1 yielded F1 scores of 0.5 and 0.5, respectively. For Model 2, the F1 scores were 0.62 (RMN) and 0.63 (DeepFace) For Model 3, the F1 scores were 0.73 (RMN) and 0.74 (DeepFace). When all modalities were combined for Model 4, the F1 scores yielded were 0.82 (RMN) and 0.81 (DeepFace).

3. Gender-specific stress prediction. Stress predictions from self-reported affect for Model 1 in Males and Females yielded F1 scores of 0.52 and 0.58. The addition of all modalities in Model 4 yielded F1 scores of 0.81 and 0.86. A secondary analysis comprising of RMN predicted affect in Model 1 yielded F1 scores of 0.47 and 0.55 in Males and Females. For Model 4, F1 scores are 0.8 and 0.83. Similarly, DeepFace predicted affect in Model 1 yielded F1 scores of 0.48 and 0.53 in Males and Females. For Model 4, F1 scores are 0.8 and 0.83.

6 Discussion

This study investigates the reasons for university CS students' elevated mental health challenges and their preferences for technological support to monitor their stress and emotional responses to academic tasks across varying situational and temporal contexts. Quantitative findings corroborate qualitative findings, revealing debugging and contextual features such as time of day are significantly correlated with elevated stress levels. Our secondary analysis on evaluating the reliability of AER tools is consistent with prior research, highlighting the need for more robust emotion detection techniques, specifically for use in digital mental health tools. Further, the highlighted role of context in stress prediction shows the promise of accurate stress detection through computer-assisted tools. In the following sections, we discuss opportunities and recommendations for future design of tools catered toward university student mental health and discuss the limitations of the study.

6.1 Designing for CS Student Mental Health

While CS students face the highest risk of mental health disorders among all engineering disciplines [34], to our knowledge, our work is the first to investigate *what* contributes to the elevated stress and burnout in this population. Our findings revealed debugging, lack of self-awareness of their mental states, and imposter syndrome as some of the pivotal challenges faced by CS students. The challenges of debugging in CS students align with prior literature, highlighting its emotional impact on students [5, 63, 112]. Experiences of students in a controlled large classroom setting will be different from their experiences in more in-the-wild naturalistic settings [15, 42]. Students face both positive and negative experiences during programming [16–18], which can affect their self-efficacy and academic outcomes [59, 69, 115]. Low self-efficacy can amplify feelings of inadequacy during debugging, reinforcing imposter syndrome, which is another prevalent theme in our findings. These imposter syndrome feelings, the doubt CS students had in their abilities after being burned out from academic tasks, despite their motivation and past successes as CS students, are consistent with prior research [91]. The interplay between self-efficacy and imposter syndrome further exacerbates their mental health issues [44, 82]. Further, the imposter syndrome feelings are more prominent in women and may deter them and other underrepresented groups from computing, with the potential to create a backslide in representation in the future of information work [31, 68].

Despite several tools being developed and tested in college students, there exists a gap in the adoption of digital mental health because there is a mismatch in tool design to their everyday experiences [50, 65, 66]. Further, these tools are not designed from students' perspectives, nor designed for students' own usage [110]. This limitation presents a unique opportunity which our study addresses by placing student self-determination and autonomy at the heart of our tool [56, 90]. Participants expressed openness to a tool that can be deployed on their computers and preferred to continuously monitor their emotions and stress to promote self-awareness. Asking participant preferences was beneficial as we gained specific insights on building a tool with the potential to help them with their mental well-being. For instance, P1 said: *Maybe some kind of symbol representing how you're feeling... that would alert you if you were getting too stressed out* and P18 said: *...if I notice that I'm in a volatile state, it might be an indication that it is time to take a five-minute break.* This indicates that incorporating features preferred by participants can support better self-management and mitigate long-term impacts of stress on their mental well-being. Therefore, we recommend that future tools aim to understand user perspectives in designing for specific populations, especially among students, to promote wider adoption in real-world settings.

6.2 Real-time Monitoring of Mental Health

Real-time tracking of stress and emotions presents a unique challenge, particularly during cognitively demanding tasks, where self-monitoring becomes inherently limited. The choice between active (e.g. ESM [64]) and passive (e.g. AER) sensing [33] to track mental well-being is incredibly challenging. Active self-tracking that utilizes digital diaries or prompts as a tool for documenting and reflecting on insights can be burdensome to users[13, 27]. On the other hand, there is a misalignment between passively sensed automated measurements and user self-reports [55]. To tackle this problem, we adopted a hybrid approach that combines algorithmic output and self-reports to promote self-awareness. Participants acknowledged the usefulness of this approach; for instance, P18 said, *"the main reason for the reduction in my stress was because often I was asked about the nature of how I was feeling and why"* and P7 stated, *"for at least people who are less conscious of their emotions, it makes sense to use facial affect recognition. I know it's not the most accurate technology, but even a little bit of info into how they seem*

to be feeling while they're working on stuff could definitely be an eye-opener for some people."

Furthermore, a contributing factor to adopting a hybrid approach was a lack of existing off-the-shelf computer-based stress detection tools. Although affect sensing is widely researched in educational settings (eg. In the classroom [25]), it is difficult to passively track stress data due to its subjective nature [51, 56]. Further, "most accurate methods rely on clinical-grade sensors and are often custom made, and expensive" [75]. The hybrid approach we adopted helped us validate our hypothesis that stress can be reliably detected using AER tools when augmented with the contextual states of the user. While we recommend systems should be designed using this hybrid approach, we also advise caution. Continuous self-tracking and reflection of mental states could be detrimental, especially among individuals with pre-existing mental health conditions. While displaying stress data can enhance users' awareness of their stress levels, it may also inadvertently amplify stress[81, 83]. Prior work highlights that users can feel overwhelmed by excessive data or experience shame when confronted with insights that reflect negatively on their well-being [56]. As systems integrate multi-modal sensing to generate inferences and individualized predictions, we recommend balancing the benefits of self-awareness with the potential risks while fostering meaningful reflection, ensuring the system supports rather than undermines users' mental well-being.

6.3 Context in Real-time Mental Health Support

The adaptability of digital mental health tools hinges on their ability to assess (a) whether the individual is in a state that requires support; (b) the specific type or amount of support required; and (c) the likelihood that the support offered will be acted upon or potentially perceived negatively [80]. This necessitates a system that dynamically adjusts to the evolving states and contexts of the person, demanding comprehensive monitoring capabilities. Our work contributes to the extensive literature on context-aware computing [2, 10, 11, 62, 95, 108, 109] by recognizing the role of context in two aspects: (i) stress prediction and (ii) self-reflection.

Researchers have determined that automated stress detection brings unique challenges because stress involves highly subjective, social, and environmental factors. Despite our findings indicating the performance of AER models was poor, the stress predictions when AER model predictions were augmented with contextual features were similar to the predictions from self-reported emotions (ground truth) augmented with contextual features. Furthermore, towards predicting momentary stress data accurately, we observed that the hour of the day (F1 score: 0.75) and type of activity (F1 score: 0.71) are important contextual factors influencing the performance of predictive models. This indicates that contextual awareness can significantly improve the accuracy of predicting stress in naturalistic settings [41, 74, 102]. Our observations support findings from prior HCI literature in the digital mental health domain; such as Bhattacharjee et al.'s work, which demonstrated that individuals' schedules and emotional states shape their responses to mHealth interventions aimed at psychological well-being [11]. Furthermore, response variability to stress among information workers performing similar tasks highlights the need for tools that consider individual-specific contexts [76]. These contextual insights are not only relevant to our target population but also extend to anyone who uses a computer as their primary medium for work, presenting an opportunity to incorporate contextual factors in informing stress predictions across diverse professional environments.

Reflection is a common design goal in HCI systems, with visualization serving as a primary medium [9, 67]. The goal of self-reflection is to influence future behaviors by providing actionable insights[47]. PI systems, sensitive to interpersonal contexts, should prioritize personalized retrospection rather than simply presenting system outputs [79]. Incorporating contextual data allows PI systems to not only facilitate self-reflection but also explain why and when specific mental states occurred. However, many existing PI systems overlook dynamic contexts during the reflection phase, missing opportunities to provide deeper insights. Our work builds on systems like the MindScope app, which uses stress prediction explanations to help users reconstruct past stressful events [58]. We expand on this work by incorporating mental state contexts, enabling participants to analyze stressful events more comprehensively. Participants in our study found the inclusion of temporal and situational context in visualizations particularly valuable for identifying patterns and triggers in their mental states. Therefore, we recommend that future PI systems dynamically integrate context into visualizations, providing clear, actionable insights tailored to individual experiences [89]. Contextual information is crucial not only for stress tracking and personal informatics but also for advancing mental well-being technologies like automated emotion recognition, as we discuss next.

6.4 The Reliability of AER Models

In our analysis, we saw that the alignment of state-of-the-art AER predictions with the ground truth (i.e., objective self-reports from users) is low. Specifically, the accuracies of Positive and Negative affect are low from RMN and DeepFace, respectively. The majority of the accuracy comes from the Neutral class. Our analysis revealed a misalignment between state-of-the-art AER predictions and ground truth self-reports, with the majority of accuracy concentrated in the Neutral class while Positive and Negative affect were less accurately predicted. This shortcoming presents a critical opportunity to improve AER performance for non-neutral states, as they are vital for mental health applications. We also observed sex-based discrepancies in AER accuracy, with higher accuracy for Female users. This may reflect the underrepresentation of non-Female users in training datasets, which aligns with prior findings that model performance is often skewed by imbalanced data. Interestingly, our results contradict earlier research suggesting Males express anger more frequently than Females [21, 38], further pointing to potential biases in the datasets used to train these models. These imbalances present opportunities to develop more inclusive and equitable systems.

Additionally, people exhibit the same emotion in different ways depending on both the internal (e.g., thoughts from the past) and external context (e.g., location, environment) [6]. For instance, the context in which users experience Positive or Negative affect—such as location (e.g., home, school, library) or activity type—varies significantly. We recommend incorporating such contextual information into AER training datasets to enhance the models' ability to

capture these nuances and improve the credibility of their predictions. Additionally, we also recommend augmenting datasets with activity-specific emotional baselines to help personalize predictions, tailoring them to individual patterns of affect. By addressing biases in gender representation and embedding diverse contextual features, AER systems can become more reliable and actionable tools for stress prediction and emotional well-being. Users aiming to incorporate AER in digital mental health tools should not rely solely on these models for emotion prediction. AER systems need to be thoroughly validated within their intended context [55] and supplemented with more reliable forms of emotion recognition, such as self-reports or physiological measures, to ensure accuracy and trustworthiness.

6.5 Ethical Considerations

Working with student participants introduces potential biases and inequities, necessitating a focus on autonomy and self-determination in digital tracking approaches [56, 90]. Privacy, transparency, and data autonomy are critical concerns, particularly when managing sensitive participant data, as emphasized by our participants [110]. To address these concerns, EmotionStream did not collect video or personally identifiable information, and all data used for analysis were anonymized. Data were stored locally, with participants maintaining full control, including the option to stop data collection or withdraw from the study at any time. Participants voluntarily shared their data at the study's conclusion and were fully informed about the nature, storage, and usage of the data collected. These measures likely alleviated some participant concerns, as evidenced by high tool engagement. However, the low accuracy of real-time AER models and the inherently subjective nature of stress monitoring present significant challenges. Participants highlighted the dual-edged nature of monitoring, emphasizing the importance of careful interpretation and action on their data [116]. We concur, advocating for future automated mental health monitoring systems to undergo rigorous testing within their intended deployment contexts to mitigate potential risks and ensure ethical use [7, 55].

6.6 Limitations and Future Work

A key limitation of this study is the need for a larger and more diverse participant pool to thoroughly evaluate potential biases. Our sample included only one participant from underrepresented minority groups, which restricts our ability to generalize challenges and preferences across diverse populations [106]. Furthermore, given that AER models are likely to exhibit variability in affect detection when tested on heterogeneous groups, this study does not account for such variations [96]. While our focus was on assessing the acceptability of the tool, future efficacy studies will prioritize diverse recruitment, aiming to include participants proportional to U.S. Census demographics. Another limitation stems from the naturalistic context in which data were collected. While capturing stress in real-world settings enhances ecological validity, it introduces uncontrolled factors such as variations in camera quality and background lighting, potentially influencing affect detection accuracy [93]. Additionally, our participant pool only included self-reported binary gender data (male and female), limiting the generalizability of our findings to a broader spectrum of gender identities. The study also did not track long-term associations between stress and contextual cues, which could offer valuable insights into the longitudinal digital phenotypes of computer science students during academic tasks. Incorporating factors such as location and weather in future research may yield a deeper understanding of their influence on affect and stress levels [14]. Lastly, the use of a Windows OS-specific application restricts scalability and limits the applicability of the tool across diverse platforms, presenting challenges for broader adoption.

7 Conclusion

This study examines the unique mental health challenges faced by CS students and their preferences for technological solutions to facilitate self-reflection through a needfinding study. In response, we developed and evaluated EmotionStream, a computer-based PI tool that integrates contextual and emotional cues to support self-reflection. Our evaluation of EmotionStream in a naturalistic setting demonstrated its acceptability among CS students and confirmed the influence of situational and temporal contexts on stress levels. Notably, stress was heightened during debugging tasks and late-night activities, aligning with our qualitative findings. This study contributes to the growing literature on mental health in the CS student population by offering actionable insights for designing PI tools that support mental well-being. EmotionStream serves as a promising starting point for designing scalable, context-aware mental health tools that cater to the specific needs of CS students.

Acknowledgments

The authors would like to thank the Distributed Research Experiences for Undergraduates program led by the CRA-Widening Participation (CRA-WP) committee and funded by the U.S. National Science Foundation through which London Bielicke was able to get involved in and contribute to this research. We also thank the reviewers for their valuable time and feedback, significantly improving the paper's quality.

References

[1] Phil Adams, Mashfiqui Rabbi, Tauhidur Rahman, Mark Matthews, Amy Voida, Geri Gay, Tanzeem Choudhury, and Stephen Voida. 2014. Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. 72–79.

[2] Jennfier Akullian, Adam Blank, Lauren Bricker, Linda DuHadway, and Christian Murphy. 2020. Supporting mental health in computer science students and professionals. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 958–959.

[3] David M Almeida, Elaine Wethington, and Ronald C Kessler. 2002. The daily inventory of stressful events: An interview-based approach for measuring daily stressors. *Assessment* 9, 1 (2002), 41–55.

[4] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. 2020. Activity river: Visualizing planned and logged personal activities for reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–9.

[5] Zahra Atiq and Michael C. Loui. 2022. A Qualitative Study of Emotions Experienced by First-year Engineering Students during Programming Tasks. *ACM Trans. Comput. Educ.* 22, 3, Article 32 (June 2022), 26 pages. doi:10.1145/3507696

[6] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. doi:10.1177/1529100619832930 arXiv:https://doi.org/10.1177/1529100619832930 PMID: 31313636.

[7] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[8] Lucy Zhang Bencharit, Yuen Wan Ho, Helene H Fung, Dannii Y Yeung, Nicole M Stephens, Rainer Romero-Canyas, and Jeanne L Tsai. 2019. Should job applicants be excited or calm? The role of culture and ideal affect in employment settings. *Emotion* 19, 3 (2019), 377.

[9] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[10] Ananya Bhattacharjee, Jiyau Pang, Angelina Liu, Alex Mariakakis, and Joseph Jay Williams. 2023. Design implications for one-way text messaging services that support psychological wellbeing. *ACM Transactions on Computer-Human Interaction* 30, 3 (2023), 1–29.

[11] Ananya Bhattacharjee, Joseph Jay Williams, Jonah Meyerhoff, Harsh Kumar, Alex Mariakakis, and Rachel Kornfield. 2023. Investigating the role of context in the delivery of text messages for supporting psychological wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[12] G Bhuvaneswari and E Emiline Joy. 2021. Assess the level of stress, sleep disturbance, depression with nomophobia among undergraduate students. *International Journal of Advanced Psychiatric Nursing* 3, 2 (2021), 24–27.

[13] Johnna Blair, Yuhan Luo, Ning F Ma, Sooyeon Lee, and Eun Kyoung Choe. 2018. OneNote Meal: A photo-based diary study for reflective meal tracking. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 252.

[14] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. 2014. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*. 477–486.

[15] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms.. In *IJCAI*, Vol. 16. 4125–4129.

[16] Nigel Bosch and Sidney D'Mello. 2014. It takes two: momentary co-occurrence of affective states during computerized learning. In *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*. Springer, 638–639.

[17] Nigel Bosch and Sidney D'Mello. 2017. The affective experience of novice computer programmers. *International journal of artificial intelligence in education* 27 (2017), 181–206.

[18] Nigel Bosch, Sidney D'Mello, and Caitlin Mills. 2013. What emotions do novices experience during their first computer programming learning session?. In *International Conference on Artificial Intelligence in Education*. Springer, 11–20.

[19] Joel Bothello and Thomas J Roulet. 2019. The imposter syndrome, or the misrepresentation of self in academic life. *Journal of Management Studies* 56, 4 (2019), 854–861.

[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[21] Leslie R Brody and Judith A Hall. 2008. Gender and emotion in context. *Handbook of emotions* 3 (2008), 395–408.

[22] Stephen L Brown, Brandye D Nobiling, James Teufel, and David A Birch. 2011. Are kids too busy? Early adolescents' perceptions of discretionary activities, overscheduling, and stress. *Journal of school health* 81, 9 (2011), 574–580.

[23] Olivia Calancie, Lexi Ewing, Laura D Narducci, Salinda Horgan, and Sarosh Khalid-Khan. 2017. Exploring how social networking sites impact youth with anxiety: A qualitative study of Facebook stressors among adolescents with an anxiety disorder diagnosis. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 11, 4 (2017).

[24] Clara Caldeira, Yu Chen, Lesley Chan, Vivian Pham, Yunan Chen, and Kai Zheng. 2017. Mobile apps for mood tracking: an analysis of features and user reviews. In *AMIA annual symposium proceedings*, Vol. 2017. American Medical Informatics Association, 495.

[25] Tara Francis Chan. 2018. A school in China is monitoring students with facial-recognition technology that scans the classroom every 30 seconds. *Business Insider* (2018).

[26] Igor Chirikov, Krista M Soria, Bonnie Horgos, and Daniel Jones-White. 2020. Undergraduate and graduate students' mental health during the COVID-19 pandemic. (2020).

[27] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. 2017. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing* 16, 1 (2017), 74–84.

[28] George P Chrousos. 2009. Stress and disorders of the stress system. *Nature reviews endocrinology* 5, 7 (2009), 374–381.

[29] Joanne Wai Yee Chung, Henry Chi Fuk So, Marcy Ming Tak Choi, Vincent Chun Man Yan, and Thomas Kwok Shing Wong. 2021. Artificial Intelligence in education: Using heart rate variability (HRV) as a biomarker to assess emotions objectively. *Computers and Education: Artificial Intelligence* 2 (2021), 100011.

[30] Tamara Cibrian-Llanderal, Montserrat Melgarejo-Gutierrez, and Daniel Hernandez-Baltazar. 2018. Stress and cognition: Psychological basis and support resources. *Health and academic achievement* 11, 10.5772 (2018).

[31] Pauline Rose Clance and Suzanne Ament Imes. 1978. The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, research & practice* 15, 3 (1978), 241.

[32] Stanley Coren and James A Russell. 1992. The relative dominance of different facial expressions of emotion under conditions of perceptual ambiguity. *Cognition and Emotion* 6, 5 (1992), 339–356.

[33] Victor P Cornet and Richard J Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics* 77 (2018), 120–132.

[34] Andrew Danowitz and Kacey Beddoes. 2018. Characterizing mental health and wellness in students across engineering disciplines. In *2018 The Collaborative Network for Engineering and Computing Diversity Conference Proceedings*.

[35] Xianghua Ding, Shuhan Wei, Xinning Gui, Ning Gu, and Peng Zhang. 2021. Data engagement reconsidered: a study of automatic stress tracking technology in use. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[36] C Emily Durbin and Sylia Wilson. 2012. Convergent validity of and bias in maternal reports of child emotion. *Psychological assessment* 24, 3 (2012), 647.

[37] Pedro Guillermo Feijóo-García, Chase Wrenn, Jacob Stuart, Alexandre Gomes De Siqueira, and Benjamin Lok. 2023. Participatory Design of Virtual Humans for Mental Health Support Among North American Computer Science Students: Voice, Appearance, and the Similarity-attraction Effect. *ACM Transactions on Applied Perception* 20, 3 (2023), 1–27.

[38] Agneta H Fischer. 1993. Sex differences in emotionality: Fact or stereotype? *Feminism & Psychology* 3, 3 (1993), 303–318.

[39] Kurara Fukumoto, Tsutomu Terada, and Masahiko Tsukamoto. 2013. A smile/laughter recognition mechanism for smile-based life logging. In *Proceedings of the 4th Augmented Human International Conference*. 213–220.

[40] Robert P Gallagher. 2015. National survey of college counseling centers 2014. (2015).

[41] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous stress detection using a wrist device: in laboratory and real life. In *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 1185–1193.

[42] Thomas Goetz, Ulrike E Nett, Sarah E Martiny, Nathan C Hall, Reinhard Pekrun, Swantje Dettmers, and Ulrich Trautwein. 2012. Students' emotions during homework: Structures, self-concept antecedents, and achievement outcomes. *Learning and Individual Differences* 22, 2 (2012), 225–234.

[43] Peter Graw, Kurt Kräuchi, Anna Wirz-Justice, and Walter Pöldinger. 1991. Diurnal variation of symptoms in seasonal affective disorder. *Psychiatry research* 37, 1 (1991), 105–111.

[44] Amirhossein Haghighi and Akram Ghorbali. 2022. The Relationship between Academic Self-Concept and Academic Self-Efficacy in College Students: The Mediating Role of Imposter Syndrome. *Contemporary Psychology, Biannual Journal of the Iranian Psychological Association* 17, 2 (2022), 7–20.

[45] Aleesha Hamid, Rabiah Arshad, and Suleman Shahid. 2022. What are you thinking?: Using CBT and Storytelling to Improve Mental Health Among College Students. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[46] Andree Hartanto, Kristine YX Lee, Yi Jing Chua, Frosch YX Quek, and Nadyanna M Majeed. 2023. Smartphone use and daily cognitive failures: A critical examination using a daily diary approach with objective smartphone measures. *British Journal of Psychology* 114, 1 (2023), 70–85.

[47] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What does all this data mean for my future mood? Actionable analytics and targeted reflection for emotional well-being. *Human–Computer Interaction* 32, 5-6 (2017), 208–267.

[48] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On being told how we feel: how algorithmic sensor feedback influences emotion perception. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–31.

[49] Noura Howell, John Chuang, Abigail De Kosnik, Greg Niemeyer, and Kimiko Ryokai. 2018. Emotional biosensing: Exploring critical alternatives. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[50] Suzanne S Hudd, Jennifer Dumlao, Diane Erdmann-Sager, Daniel Murray, Emily Phan, Nicholas Soukas, and Nori Yokozuka. 2000. Stress at college: effects on health habits, health status and self-esteem. *College student journal* 34, 2 (2000).

[51] Stephen Hutt, Joseph F Grafsgaard, and Sidney K D'Mello. 2019. Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[52] Jie Ji, Christian Murphy, Brianna Blaser, and Jennifer Akullian. 2024. Experiences of Undergraduate Computer Science Students Living with Mental Health Conditions. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 597–603.

[53] Matthew Jörke, Yasaman S Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. 2023. Pearl: A technology probe for machine-assisted reflection on personal data. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 902–918.

[54] Charalampos Karyotis, Faiyaz Doctor, Rahat Iqbal, Anne E James, and Victor Chang. 2016. A Fuzzy Modelling Approach of Emotion for Affective Computing Systems.. In *IoTBD*. 453–460.

[55] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. 2022. "I didn't know I looked angry": Characterizing observed emotion and reported affect at work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[56] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for mental wellness: understanding expert perspectives and student experiences. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 629–641.

[57] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, et al. 2022. Prediction for retrospection: Integrating algorithmic stress prediction into personal informatics systems for college students' mental health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.

[58] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 279, 20 pages. doi:10.1145/3491102.3517701

[59] Päivi Kinnunen and Beth Simon. 2012. My program is ok–am I? Computing freshmen's experiences of doing programming assignments. *Computer Science Education* 22, 1 (2012), 1–28.

[60] Patrick Klaiber, Jin H Wen, Anita DeLongis, and Nancy L Sin. 2021. The ups and downs of daily life during COVID-19: Age differences in affect, stress, and positive events. *The Journals of Gerontology: Series B* 76, 2 (2021), e30–e37.

[61] Nobuyoshi Komuro, Tomoki Hashiguchi, Keita Hirai, and Makoto Ichikawa. 2021. Predicting individual emotion from perception-based non-contact sensor big data. *Scientific Reports* 11 (01 2021), 2317. doi:10.1038/s41598-021-81958-2

[62] Rachel Kornfield, Renwen Zhang, Jennifer Nicholas, Stephen M Schueller, Scott A Cambo, David C Mohr, and Madhu Reddy. 2020. " Energy is a Finite Resource": Designing Technology to Support Individuals across Fluctuating Symptoms of Depression. In *Proceedings of the 2020 CHI Conference on Human factors in Computing systems*. 1–17.

[63] Essi Lahtinen, Kirsti Ala-Mutka, and Hannu-Matti Järvinen. 2005. A study of the difficulties of novice programmers. In *Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* (Caparica, Portugal) *(ITiCSE '05)*. Association for Computing Machinery, New York, NY, USA, 14–18. doi:10.1145/1067445.1067453

[64] Reed Larson and Mihaly Csikszentmihalyi. 2014. *The Experience Sampling Method*. Springer Netherlands, Dordrecht, 21–34. doi:10.1007/978-94-017-9088-8_2

[65] Emily G Lattie, Rachel Kornfield, Kathryn E Ringland, Renwen Zhang, Nathan Winquist, and Madhu Reddy. 2020. Designing mental health technologies that support the social ecosystem of college students. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[66] Kwangyoung Lee and Hwajung Hong. 2018. MindNavigator: Exploring the stress and self-interventions for mental wellness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[67] 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.

[68] Danielle Lindemann, Dana Britton, and Elaine Zundl. 2016. "I don't know why they make it so hard here": Institutional factors and undergraduate women's STEM participation. *International Journal of Gender, Science and Technology* 8, 2 (2016), 221–241.

[69] Alex Lishinski, Aman Yadav, and Richard Enbody. 2017. Students' emotional reactions to programming projects in introduction to programming: Measurement approach and influence on learning outcomes. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*. 30–38.

[70] Carolyn M Mazure. 1998. Life stressors as risk factors in depression. *Clinical Psychology: Science and Practice* 5, 3 (1998), 291.

[71] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 849–858.

[72] Jennifer Melcher, Ryan Hays, and John Torous. 2020. Digital phenotyping for mental health of college students: a clinical review. *BMJ Ment Health* 23, 4 (2020), 161–166.

[73] Jennifer Melcher, Joel Lavoie, Ryan Hays, Ryan D'Mello, Natali Rauseo-Ricupero, Erica Camacho, Elena Rodriguez-Villa, Hannah Wisniewski, Sarah Lagan, Aditya Vaidyam, et al. 2023. Digital phenotyping of student mental health during COVID-19: an observational study of 100 college students. *Journal of American College Health* 71, 3 (2023), 736–748.

[74] Varun Mishra, Tian Hao, Si Sun, Kimberly N Walter, Marion J Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the role of context in perceived stress detection in the wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1708–1716.

[75] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2020. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Trans. Comput. Healthcare* 1, 2, Article 8 (April 2020), 30 pages. doi:10.1145/3361562

[76] Mehrab Bin Morshed, Javier Hernandez, Daniel McDuff, Jina Suh, Esther Howe, Kael Rowan, Marah Abdin, Gonzalo Ramos, Tracy Tran, and Mary Czerwinski. 2022. Advancing the Understanding and Measurement of Workplace Stress in Remote Information Workers from Passive Sensors and Behavioral Data. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. doi:10.1109/ACII55700.2022.9953824

[77] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. 2019. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21.

[78] Carol T Mowbray, James M Mandiberg, Catherine H Stein, Sandra Kopels, Caroline Curlin, Deborah Megivern, Shari Strauss, Kim Collins, and Robin Lett. 2006. Campus mental health services: Recommendations for change. *American Journal of Orthopsychiatry* 76, 2 (2006), 226–237.

[79] Elizabeth L Murnane, Tara G Walker, Beck Tench, Stephen Voida, and Jaime Snyder. 2018. Personal informatics in interpersonal contexts: towards the design of technology that supports the social ecologies of long-term mental health management. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[80] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* (2018), 1–17.

[81] Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M Almeida, and Santosh Kumar. 2024. Momentary Stressor Logging and Reflective Visualizations: Implications for Stress Management with Wearables. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[82] Csilla Pákozdy, Jemima Askew, Jessica Dyer, Phoebe Gately, Leya Martin, Kenneth I Mavor, and Gillian R Brown. 2024. The imposter phenomenon and its relationship with self-efficacy, perfectionism and happiness in university students. *Current Psychology* 43, 6 (2024), 5153–5162.

[83] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with stress through pop-culture. In *Proceedings of the 8th international conference on pervasive computing technologies for healthcare*. 109–117.

[84] Paola Pedrelli, Maren Nyer, Albert Yeung, Courtney Zulauf, and Timothy Wilens. 2015. College students: mental health problems and treatment considerations. *Academic psychiatry* 39 (2015), 503–511.

[85] Luan Pham. [n. d.]. The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 4513–4519.

[86] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[87] Stephanie Pinder-Amaker and Catherine Bell. 2012. A bioecological systems approach for navigating the college mental health crisis. *Harvard Review of Psychiatry* 20, 4 (2012), 174–188.

[88] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 281–290.

[89] Amon Rapp and Federica Cena. 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94 (2016), 1–17.

[90] John Rooksby, Alistair Morrison, and Dave Murray-Rust. 2019. Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[91] Adam Rosenstein, Aishma Raghu, and Leo Porter. 2020. Identifying the prevalence of the impostor phenomenon among computer science students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 30–36.

[92] Samara Ruiz, Sven Charleer, Maite Urretavizcaya, Joris Klerkx, Isabel Fernández-Castro, and Erik Duval. 2016. Supporting learning by considering emotions: tracking and visualization a case study. In *Proceedings of the sixth international conference on learning analytics & knowledge*. 254–263.

[93] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* 19, 8 (2019), 1863.

[94] Pedro Sanches, Kristina Höök, Elsa Vaara, Claus Weymann, Markus Bylund, Pedro Ferreira, Nathalie Peira, and Marie Sjölinder. 2010. Mind the body! Designing a mobile stress management application encouraging personal reflection. In *Proceedings of the 8th ACM conference on designing interactive systems*. 47–56.

[95] Matthew Saponaro, Ajith Vemuri, Greg Dominick, and Keith Decker. 2021. Contextualization and individualization for just-in-time adaptive interventions to reduce sedentary behavior. In *Proceedings of the conference on health, inference, and learning*. 246–256.

[96] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[97] Sefik Ilkin Serengil and Alper Ozpinar. 2020. Lightface: A hybrid deep face recognition framework. In *2020 innovations in intelligent systems and applications conference (ASYU)*. IEEE, 1–5.

[98] Sefik Ilkin Serengil and Alper Ozpinar. 2021. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 1–4.

[99] Sefik Ilkin Serengil and Alper Ozpinar. 2023. An Evaluation of SQL and NoSQL Databases for Facial Recognition Pipelines. (2023).

[100] Astha Sharma and Shaun Canavan. 2021. Multimodal Physiological-Based Emotion Recognition. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*. Springer-Verlag, Berlin, Heidelberg, 101–113. doi:10.1007/978-3-030-68790-8_9

[101] Derrick Silove, Ingrid Sinnerbrink, Annette Field, Vijaya Manicavasagar, and Zachary Steel. 1997. Anxiety, depression and PTSD in asylum-seekers: associations with pre-migration trauma and post-migration stressors. *The British Journal of Psychiatry* 170 (1997), 351.

[102] Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, Walter De Raedt, Jan Cornelis, Olivier Janssens, Sofie Van Hoecke, Stephan Claes, et al. 2018. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* 1, 1 (2018), 67.

[103] Lígia Maria Soares Passos, Christian Murphy, Rita Zhen Chen, Marcos Gonçalves de Santana, and Giselle Soares Passos. 2020. The prevalence of anxiety and depression symptoms among brazilian computer science students. In *Proceedings of the 51st acm technical symposium on computer science education*. 316–322.

[104] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*. 715–724.

[105] Marijn Ten Thij, Krishna Bathina, Lauren A Rutter, Lorenzo Lorenzo-Luaces, Ingrid A van de Leemput, Marten Scheffer, and Johan Bollen. 2020. Depression alters the circadian pattern of online activity. *Scientific reports* 10, 1 (2020), 17272.

[106] Adela C Timmons, Jacqueline B Duong, Natalia Simo Fiallo, Theodore Lee, Huong Phuc Quynh Vo, Matthew W Ahle, Jonathan S Comer, LaPrincess C Brewer, Stacy L Frazier, and Theodora Chaspari. 2023. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science* 18, 5 (2023), 1062–1096.

[107] Elexandra Tran, Liuming Huang, Michelle Craig, and Sadia Sharmin. 2022. The Impact of Gratitude Journaling on CS1 Students. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2*. 52–52.

[108] Ajith Vemuri, Keith Decker, Mathew Saponaro, and Gregory Dominick. 2021. Multi agent architecture for automated health coaching. *Journal of medical systems* 45, 11 (2021), 95.

[109] Ajith Vemuri, Megan Heintzelman, Alex Waad, Matthew Louis Mauriello, Keith Decker, and Gregory Dominick. 2023. Towards dynamic action planning with user preferences in automated health coaching. *Smart Health* 28 (2023), 100389.

[110] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz, and Betsy Disalvo. 2020. Sensing Affect to Empower Students: Learner Perspectives on Affect-Sensitive Technology in Large Educational Contexts. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 63–76. doi:10.1145/3386527.3405917

[111] 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.

[112] Jacqueline Whalley, Amber Settle, and Andrew Luxton-Reilly. 2021. Novice Reflections on Debugging. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 73–79. doi:10.1145/3408877.3432374

[113] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 506–523.

[114] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing* (2021).

[115] Stephanie Yang, Miles Baird, Eleanor O'Rourke, Karen Brennan, and Bertrand Schneider. 2024. Decoding Debugging Instruction: A Systematic Literature Review of Debugging Interventions. *ACM Trans. Comput. Educ.* 24, 4, Article 45 (Nov. 2024), 44 pages. doi:10.1145/3690652

[116] Shoshana Zuboff. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology* 30, 1 (2015), 75–89.