Towards Human-Centered Behavioral Sensing for Student Support

Han Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Jennifer Mankoff, Co-chair
Anind K. Dey, Co-chair
Yin Tat Lee

Program Authorized to Offer Degree:
Department of Computer Science & Engineering

University of Washington

Abstract

Towards Human-Centered Behavioral Sensing for Student Support

Han Zhang

Co-Chairs of the Supervisory Committee:

Jennifer Mankoff
Department of Computer Science & Engineering

Anind K. Dey
Information School

Behavioral sensing technologies hold significant potential to enhance human support systems, especially in high-stakes domains such as education and mental well-being. However, existing research often prioritizes model performance, limiting its ability to address real-world needs, adapt to diverse populations, or surface potential risks. Focusing on student support, this dissertation advances a human-centered approach to behavioral sensing by deepening the understanding of students' needs, examining potential harms embedded in current practices, and developing behavioral models that align with Human-Centered Machine Learning (HCML) principles.

This work begins with a six-year longitudinal study that combines passive sensing and self-reported data to capture the everyday academic and mental well-being experiences of college students. This is followed by a mixed-method study examining how the transition to online learning during COVID-19—a major life event that disrupted routines, support systems, and access to resources—impacted students with disabilities and mental health concerns—revealing their needs from the onset of the pandemic through the following academic year.

These empirical findings motivate a deeper investigation into the ethical risks and fairness concerns surrounding behavioral sensing in real-world deployment. Through both quantitative and qualitative studies, we provide evidence of algorithmic bias embedded in existing behavioral models and uncover broader ethical challenges across the behavioral sensing lifecycle. This work

highlights the unique nature of fairness in behavioral sensing, identifies stage-specific vulnerabilities, and surfaces systemic barriers to fair and accountable system design. Drawing on these insights, we propose a reflexive fairness framework and actionable guidelines to support more ethically aligned development and evaluation practices.

Informed by these insights, we develop and evaluate three predictive modeling approaches that identify at-risk students as early as the first week of an academic term. These approaches integrate fairness, interpretability, and generalizability as core design objectives. Our findings demonstrate the feasibility of operationalizing HCML in behavioral modeling, while also highlighting key trade-offs and design tensions that emerge in practice.

Finally, we discuss open challenges and future directions for building responsible behavioral sensing systems that are not only technically robust but also ethically grounded, inclusive, and attuned to the lived realities of those they aim to support.

Acknowledgments

Reflecting on the past few years, completing a PhD during a global pandemic brought moments of challenge, uncertainty, and reflection—but also unexpected connection, resilience, and growth. I could not have made it through without the many people who stood beside me, believed in me, and offered their support in ways big and small. I am incredibly grateful to the mentors, collaborators, friends, and family who helped me grow along the way.

First, I would like to thank my advisors, Jennifer Mankoff and Anind Dey, for helping me grow into a more mature researcher and person. I am deeply grateful that they chose to admit me as a PhD student—even though my computer science background at the time could generously be described as "aspirational". Over the past five years, both of their expertise and ways of thinking have shaped how I approach research—and how I see the world. Jen introduced me to the field of accessibility and continually pushed me to think beyond the technical: to ask not just what we can build, but who it is for, who might be left out, and how our work can make a meaningful difference. Anind opened the door to behavioral sensing and well-being, and always reminded me to step back and see the bigger picture and real-world impact. I feel incredibly lucky to have had two advisors who brought such complementary wisdom, values, and support to this journey. Beyond research, I am also grateful for the care they showed toward my mental well-being and work-life balance—they never let me sacrifice my health for a paper deadline, and constantly reminded me that we are people first, not just researchers.

I want to express my gratitude to my thesis committee members, Gary Hsieh and Yin Tat Lee. Gary has devoted a great deal of time and effort to supporting my progress since my General Examination. His thoughtful feedback and suggestions have strengthened not only my dissertation, but also the way I think about research communication and positioning. Yin Tat brought a refreshing perspective to my work, and I have always appreciated his sharp questions,

clear thinking, and generosity in engaging with ideas outside his core area. I would also like to thank several UW faculty members who have supported me during this journey. Jon Froehlich has modeled what it means to be passionate about research and mentorship, and his kind words of encouragement before each major milestone always meant a great deal. James Fogarty once sent me a kind and thoughtful email after one of my public talks—a talk during which I more or less froze on stage. His empathy in that moment, and his later advice on job searching and setting priorities, have stayed with me ever since.

I am very fortunate to have been mentored by and collaborated with many generous and brilliant researchers. Yasaman Sefidgar quite literally taught me how to code when I first entered the field. Paula Nurius warmly welcomed me into perspectives from psychology and social work. Sally Goldman dedicated a quarter to supporting my career development. Koustuv Saha encouraged me to be bold and try new things, and he has been a "live dictionary" whenever I needed references. Vedant Das Swain, with his sharp insights and even sharper black humor, made research conversations both rich and entertaining. Abdelkareem Bedri and Gierad Laput provided an incredibly rewarding internship at Apple, with more than enough technical resources and support to pursue my work. Colin Lea reminded me of the importance of defining my own research path rather than simply following trends. I am also grateful to Vasileios Baltatzis, Jennifer Brown, Kaiming Cheng, Olivia Figueira, Leah Findlater, Nan Gao, Kevin Kuehn, Raja Kushalnagar, Jaewook Lee, Avery Mack, Margaret Morris, Lorna Quandt, Yiyi Ren, Eve Riskin, Flora Salim, Yilun Sheng, Rotem Shalev-Arkushin, Katie Malloy Spink, Xia Su, Leijie Wang, and Xuhai Xu for the many thoughtful and fruitful discussions.

Of course, no amount of academic support would have sustained me without the friendships that brought joy, comfort, and perspective to this journey. Thank you to Katherine Anderson, Sarah Heath, Dan Sears, and Peter Hill, who made our arrival in the U.S. feel far less daunting. From helping us settle in to sharing meals, laughs, and long conversations, you made a new country feel like home. Thank you to Venkatesh Potluri, Sudheesh Singanamalla, Kate Glazko, and Priyal Suneja, who never hesitated to drag me out of my stress bubble and stand by my side when I was upset. I believe there are still more Disney trips ahead—more to scream through,

and more joy to share. Thank you to Chu Li, Danli Luo, Yuhao Wan, Chenxi Yang, Haiyan He, Anqi Gao, Yizhong Wang, Xiangfeng Zhu, Dong He, Xieyang Xu, Karan Ahuja, Chien-Yu Lin, Linxing (Preston) Jiang, Hancheng Cao, and Mingxin Gu for the many hikes, games, and hotpot dinners that made my life fuller and brighter. And thank you to Tongshuang (Sherry) Wu and Michael Xieyang Liu for your companionship, shared meals, and endless morning coffees during my time in Pittsburgh.

Without my family, I would never have come this far. I would like to thank my parents, Wujun Zhang and Yanli Chen, for their unconditional love and unwavering support. I selfishly decided to go to Singapore in 2016 and then to the U.S. in 2018, chasing dreams that took me farther and farther from home. Even though it meant missed holidays, time zone mismatches, and long stretches apart, you never once questioned my choices—only reminded me to eat well, sleep enough, and keep going. Your quiet strength and endless belief in me have carried me through more than you know. I hope this milestone makes you proud. I also want to thank my grandparents, who always found ways to spoil me with snacks and a little extra money for the things I wanted—making me feel like the happiest kid in the world. I hope, wherever you are, you can see where I am today.

Last but not least, I want to thank my husband, Haotian Jiang, who has been my anchor through every high and low of this journey. Thank you for believing in me even when I doubted myself, for sending handwritten cards to document and celebrate my small wins, for holding my hand as we explored this beautiful world together, for cooking all the meals that ensured I didn't accidentally starve while writing about well-being, and for standing by me through sleepless nights, looming deadlines, and moments of uncertainty. Every day, I watch and learn from you to become a better researcher and a better person. Once, I joked that without you, I would never have come to the U.S. if I had not met you—but the truth is, I am deeply grateful for everything I have gained on this journey, and I cannot imagine it without you. Thank you for walking beside me, every step of the way.

To my parents,

who gave me roots to grow and wings to fly.

To my husband,

who held my hands through every storm and every sunrise.

Contents

List of Tables

List of Figures

List of Acronyms

1D-CNN	One-Dimensional Convolutional Neural Network
AI	Artificial Intelligence
B-H	Benjamini-Hochberg
CSCW	Computer-Supported Cooperative Work & Social Computing
CV	Computer Vision
EMAs	Ecological Momentary Assessments
FDR	False Discovery Rate
FNR	False Negative Rate
FPR	False Positive Rate
FPs	False Positives
GPA	Grade Point Average
HCI	Human-Computer Interaction
HCML	Human-Centered Machine Learning
IRB	Institutional Review Board
LMS	Learning Management Systems
LOPO-CV	Leave-One-Participants-Out Cross-Validation
LOSO-CV	Leave-One-Subject-Out Cross-Validation
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MOOC	Massive Open Online Courses
MTL	Multi-Task Learning
NLP	Natural Language Processing

OLS	Online Learning Systems
POTS	Postural Orthostatic Tachycardia Syndrome
SVM	Support Vector Machine
TA	Thematic Analysis
TPR	True Positive Rate
TPs	True Positives
Ubicomp	Ubiquitous Computing
UW	University of Washington
XAI	Explainable AI

Chapter 1

Introduction

In the past decade, data, Machine Learning (ML), and Artificial Intelligence (AI) have rapidly evolved, bringing with them the promise—and the challenge—of reshaping how people live, learn, and connect. This transformation has significantly expanded the possibilities for understanding human behavior and for designing systems that provide meaningful support in real-world contexts, including education, well-being, and accessibility [Jordan and Mitchell, 2015; Mohr et al., 2017b; Bigham and Carrington, 2018]. At the same time, it raises a pressing human-centered question that motivates this dissertation: **How can we leverage these capabilities to responsibly support diverse human needs and prevent unintended consequences from ML/AI systems developed without sufficient ethical considerations?**

Answering this question requires a deep understanding of human experience, a critical awareness of the harms and unintended consequences that can arise from data-driven systems, and the development of technologies that are context-aware and socially responsive. This, in turn, demands collecting the "right" data—data that reflects the complexity, variability, and situated nature of everyday life by capturing both observable behavioral patterns and subjective experiences through direct engagement. It also calls for close attention to how design and evaluation decisions—spanning from problem ideation to system deployment—can reinforce existing inequities or introduce new forms of harm for all communities involved, including users, researchers, and broader institutional actors. Ultimately, this means building systems that are not only technically robust, but also attuned to the environments in which they operate and accountable for the social consequences they may produce.

Traditionally, researchers have relied on surveys, interviews, and lab studies to understand hu-

man experience. While invaluable, these methods are often constrained by retrospective reporting, artificial settings, and limited temporal resolution [Shiffman et al., 2008]. Recently advances in passive *behavioral sensing*—through mobile phones, wearables, and digital platforms—have enabled researchers to capture fine-grained, continuous data about people's everyday behaviors and environments [Lane et al., 2010; Harari et al., 2016; Nepal, 2024]. Many recent studies have also combined behavioral traces with self-reports, such as Ecological Momentary Assessments (EMAs), to better capture subjective context [Saha et al., 2017; Wang et al., 2018c; Huckins et al., 2020a]. However, most of these studies remain small in scale, short in duration, and privately collected in nature—limiting their reproducibility, generalizability, and utility for understanding long-term, population-level trends. Moreover, relatively few pair behavioral and self-report data with in-depth participant engagement—such as interviews—that prioritize individuals' first-hand perspectives and interpretations. This missing layer of engagement is particularly important for studying the needs of marginalized or underrepresented groups, whose experiences may be misrepresented or overlooked in datasets that are neither inclusive nor publicly accessible.

Building on these multimodal data sources, ML/AI models have been increasingly used to detect behavioral patterns, predict risks, and infer states in domains where signals are often subtle and context-dependent, such as mental health, academic performance, and workplace well-being [Wang et al., 2015b; Adler et al., 2020; Das Swain et al., 2020]. While many of these models achieve strong predictive performance, they are commonly optimized for narrow tasks on constrained datasets, with limited regard for generalization across diverse populations and contexts. Moreover, such modeling practices frequently rely on proxy features or behavioral indicators without examining the assumptions embedded within them—raising concerns about bias, misrepresentation, or unintended consequences when models are deployed in real-world environments [Holstein et al., 2019a]. These challenges are compounded by the lack of transparency and interpretability in many behavioral models, which can undermine both the utility of behavioral models for practitioners and the trust of those affected by their outputs [Doshi-Velez and Kim, 2017; Banovic et al., 2023]. As behavioral sensing is increasingly applied in high-stakes contexts, it becomes crucial to rethink what constitutes best practice—moving beyond accuracy as the sole

benchmark and embracing more holistic approaches that surface and mitigate harms throughout the design and deployment lifecycle.

1.1 Thesis Overview and Statement

To address these challenges, this dissertation takes a human-centered approach to behavioral sensing and modeling. First, by capturing multimodal behavioral and self-report data across a six-year longitudinal study, we provide a rich, real-world foundation for examining individual behaviors and situational factors. Grounded in this data and close engagement with participants, we then offer a nuanced understanding of people's lived experiences and needs, particularly those of underrepresented populations. Second, by empirically auditing existing modeling practices and engaging directly with researchers with the field, we surface how assumptions embedded across the behavioral sensing pipeline can lead to misaligned outcomes or unintentional consequences. Third, we design behavioral models that integrate HCML principles—such as fairness, explainability, and generalizability—aiming to balance technical robustness and social alignment.

To ground this work, I focus primarily an educational setting—specifically, college student academic well-being—as a high-impact context for behavioral sensing and intervention. This setting offers a concrete lens through which to examine the challenges and opportunities of deploying behavioral technologies in real-world environments. It enables deep engagement with complex challenges such as academic struggles, mental health disparities, students' lived experiences during systemic disruptions (such as the COVID-19 pandemic), and the ethical tensions surrounding predictive modeling. Together, these efforts support the central claim of this dissertation:

> Behavioral sensing has the potential to transform how we understand and support people in domains such as education–but only if it moves beyond narrow performance goals to capture diverse lived experiences, reflect human values, and anticipate potential harms. Using college student academic well-being as a test case, I combine data from a six-year study with qualitative insights to uncover how student needs vary across backgrounds and contexts, revealing the needs for personalized approaches. Based on this understanding and

interviews with behavioral sensing experts, I developed a reflexive fairness framework that identifies stage-specific risks across the behavioral sensing lifecycle and offers actionable guidelines. I also introduced three predictive modeling approaches that balance accuracy with human values to better support early identification of at-risk students.

In the next section, I outline the structure of this dissertation.

1.2 Thesis Outline

This dissertation is structured around three interrelated lines of research:

Chapter 2: Capturing Student Experience and Understanding Their Needs. This chapter investigates how behavioral and self-report data can be used to capture and understand individual needs in everyday contexts. It begins with introducing a six-year longitudinal study aimed at gaining deeper insights into college student life, as well as a publicly available multi-year dataset derived from this study (Section 2.3). Drawing on this dataset and direct engagement with participants, we examine patterns in academic and mental well-being, with particular attention to students with disabilities and mental health concerns as they navigated the systemic disruptions caused by the COVID-19 pandemic (Section 2.4). Our findings reveal that, compared to their peers, students with disabilities expressed greater concern about online learning in the early stages of the pandemic. Follow-up interviews further illuminate both the positive and negative aspects of their online learning experiences, while underscoring the need for learning systems that are responsive to the diverse and evolving needs of students with disabilities.

Chapter 3: Investigating Harms in Current Behavioral Sensing Practice. This chapter audits current behavioral sensing practices using data from our broader study, which combines algorithmic bias evaluation experiments with interviews conducted with domain researchers across five countries, spanning both academia and industry. It (1) evaluates fairness across demographic and situational factors in nine widely used depression detection models (Section 3.4), and (2) explores field-level awareness of bias and broader harms—extending the lens from algorithmic bias to fairness considerations throughout the entire behavioral sensing

lifecycle—through qualitative inquiry (Section 3.3). The quantitative results provide empirical evidence of bias embedded in existing behavioral models, while the qualitative findings reveal how fairness is understood in practice, surface potential risks at each stage of the pipeline, and highlight structural and epistemic barriers that hinder fairness-oriented research in this domain. Drawing on these insights, the chapter concludes with practical guidelines for identifying, anticipating, and mitigating harms in behavioral sensing research.

Chapter 4: Towards Human-Centered Behavioral Modeling. This chapter explores the development of predictive behavioral models that maintain high technical performance while preserving ethical integrity. Focusing on early academic performance prediction, it introduces three modeling approaches that incorporate principles of fairness, explainability, and generalizability. This work represents one of the first efforts to integrate multiple HCML principles simultaneously within behavioral modeling. Our results demonstrate that these models achieve strong predictive performance and can identify at-risk students as early as the first week of an academic term. In addition, we outline how academic behavioral signals can inform timely and targeted support for student success. While each model prioritizes different aspects of HCML, together they offer insight into the trade-offs and design considerations involved in building more responsible and context-aware predictive systems.

In conclusion, **Chapter 5** reflects on the key contributions and insights of this dissertation and outlines promising directions for future research.

1.3 Prior Publication and Authorship

While I am the primary contributor to most of the research detailed in this dissertation, it would not have been possible without the invaluable input and collaboration of my co-authors. To reflect and acknowledge their contributions, I use the first-person plural throughout this dissertation.

The research presented in this dissertation is heavily based on the following jointly authored prior publications. For each publication, I provide its bibliographic information, indicate the chapter in which it is discussed, and include a brief summary of my contribution statement.

[1] **Han Zhang**, Margaret E. Morris, Paula S. Nurius, Kelly Mack, Jennifer Brown, Kevin

Kuehn, Yasaman S. Sefidgar, Xuhai Xu, Eve A. Riskin, Anind K. Dey, Jennifer Mankoff. Impact of Online Learning in the Context of COVID-19 on Undergraduates with Disabilities and Mental Health Concerns. *ACM Transactions on Accessible Computing*. 2022. [PDF]—In Chapter 2

Contribution statement: I co-conceptualized the research problem, conducted the data analysis, and led the writing of the paper.

[2] Xuhai Xu, **Han Zhang**, Yasaman S. Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula S. Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve A. Riskin, Jennifer Mankoff, Anind K. Dey. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generation. *Advances in Neural Information Processing Systems*. 2022. [PDF]—In Chapter 2

Contribution statement: I co-led data cleaning and curation, maintained the data codebook, assisted with data collection and quality assurance, supported analysis and benchmarking, and contributed to the paper writing and supplementary materials.

[3] **Han Zhang**, Leijie Wang, Yilun Sheng, Xuhai Xu, Jennifer Mankoff, Anind K. Dey. A Framework for Designing Fair Ubiquitous Computing Systems. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the International Symposium on Wearable Computing*. 2023. [PDF]—In Chapter 3

Contribution statement: I conceptualized the research problem, and led both the study and the writing of the paper.

[4] **Han Zhang**, Vedant Das Swain, Leijie Wang, Nan Gao, Yilun Sheng, Xuhai Xu, Flora Salim, Koustuv Saha, Anind K. Dey, Jennifer Mankoff. Illuminating the Unseen: A Framework for Designing and Mitigating Context-induced Harms in Behavioral Sensing. *arXiv preprint*. 2024. [PDF]—In Chapter 3

Contribution statement: I conceptualized the research problem, co-developed the code, conducted the analysis, and led the manuscript writing.

[5] **Han Zhang**, Yiyi Ren, Paul Nurius, Jennifer Mankoff, Anind K. Dey. Towards Human-Centered Early Prediction Models for Academic Performance in Real-World Contexts. *Ac-*

cepted to Computer-Supported Cooperative Work & Social Computing. 2025. [PDF]—In Chapter 4

Contribution statement: I conceptualized the research problem, developed the machine learning algorithms and corresponding code, conducted the data analysis and evaluation, and led the writing of the paper.

The following jointly authored publications are also referenced briefly in this dissertation.

[1] Margaret E. Morris, Kevin Kuehn, Jennifer Brown, Paula S. Nurius, **Han Zhang**, Yasaman S. Sefidgar, Xuhai Xu, Eve A. Riskin, Anind K. Dey, Sunny Consolvo, Jennifer Mankoff. College from Home During COVID-19: A Mixed-Methods Study of Heterogeneous Experiences. *PLOS ONE*. 2021. [PDF]—In Chapter 2

Contribution statement: I contributed to the data preparation for this paper.

[2] Paula S. Nurius, Yasaman S. Sefidgar, Kevin Kuehn, Jake Jung, **Han Zhang**, Olivia Figueira, Eve A. Riskin, Anind K. Dey, Jennifer Mankoff. Distress Among Undergraduates: Marginality, Stressors and Resilience Resources. *Journal of American College Health.* 2023. [PDF]—In Chapter 2

Contribution statement: I contributed to the statistic analysis for this paper.

[3] Katie M. Spink, **Han Zhang**, Paula S. Nurius, Katherine Seldin, Yiyi Ren, and Kate T. Foster. The Effects of Proximal and Distal Forms of Stress on College Student Mental Health and Affective Well-being. *Journal of American College Health.* 2025. [PDF]—In Chapter 2

Contribution statement: I contributed to the data preparation and analysis for this paper and assisted with paper writing.

[4] Xuhai Xu, Xin Liu, **Han Zhang**, Weichen Wang, Subigya Nepal, Yasaman S. Sefidgar, Woosuk Seo, Kevin Kuehn, Jeremy Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew T. Campbell, Anind K. Dey, Jennifer Mankoff. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.*

2023. [PDF]—In Chapter 3

Contribution Statement: I managed data codebook development, cleaning, and quality assurance at Institution 1, led dataset curation and visualization, assisted with analysis and validation, and contributed to writing.

Chapter 2

Capturing Student Experience and Understanding Their Needs

2.1 Overview

Supporting college students' academic success and well-being has been an area of active research across multiple disciplines such as education [Tinto, 1975; Kuh et al., 2006], computer science [Wang et al., 2015a; Nepal et al., 2022], and psychology [Nurius et al., 2015; Lattie et al., 2019]. While interest in using classroom and limited out-of-classroom data has grown, large-scale, longitudinal, and publicly available datasets that capture the multifaceted, dynamic nature of student life—particularly across diverse populations—remain rare. Moreover, few studies directly engage students to contextualize these data or examine how their needs change over time.

In the first half of this chapter, we introduce a longitudinal data collection study conducted at the University of Washington (UW) from 2018 to 2023, aimed at developing a holistic understanding of college students' daily lives, both within and beyond the classroom. This study captures diverse aspects of student life, including daily behaviors (e.g., screen time, sleep, and physical activity), mental and physical health, and academic outcomes. To enhance the representativeness of our dataset, we intentionally recruited participants from diverse backgrounds, including first-generation college students, students with disabilities, and members of underrepresented minority groups. This study complements prior work conducted at a private university [Dartmouth, 2014; Wang et al., 2014a], offering a broader lens on student life at a large public institution. To facilitate usability and reusability, we invested significant effort into data preprocessing and documentation—key bottlenecks in data science workflows. This included the development of ro-

bust data-cleaning pipelines, configuration files, and structured metadata, resulting in a dataset that is clean, well-documented, and ready for analysis. The first four years of data are publicly available on PhysioNet, with additional documentation on the GLOBEM Dataset Page and the UWEXP Study Page. We hope this effort could advance research on student well-being and enabling the development of socially informed behavioral models and interventions.

During our data collection, one major life event that significantly shaped student experiences was the unexpected outbreak of the COVID-19 pandemic. To understand its impact—particularly on vulnerable student populations—the second half of this chapter focuses on the experiences of college students with disabilities, including those with mental health conditions, in the context of the abrupt transition to online learning. We adopted a mixed-methods approach, comparing 28 undergraduate students with disabilities to their peers during 2020 to explore differences and commonalities in educational concerns, stress levels, and COVID-19–related adversities, including financial pressures. To gain deeper insight into their lived experiences, we conducted qualitative analyses of open-ended interview responses. This allowed us to examine both positive and negative aspects of online learning as experienced by students with disabilities and mental health challenges. Our goal is to illuminate how systemic disruptions interact with individual needs, and to inform more inclusive educational support practices moving forward.

2.2 Background and Related Work

A Review of Behavioral Sensing

The rapid evolution of sensing technologies has unlocked new possibilities for tracking and understanding human activities. Behavioral sensing technology, which involves using sensing to capture, model, and predict human behaviors, offers a broad spectrum of applications. This technology, in contrast to the traditional manual approach of using questionnaire-collected data for the same tasks, facilitates continuous, automated, and unobtrusive gathering of *context* [Das Swain et al., 2023; Cornet and Holden, 2018]. Here, context refers to capturing all information related to the interactions among users, applications, and their environment [Dey, 2001; Dey et al., 2001].

Behavioral Sensing for Well-Being. A substantial body of work has used behavioral sensing to detect and model psychological well-being, especially stress, anxiety, and depression [Wang et al., 2014b; Canzian and Musolesi, 2015b; Saeb et al., 2015b; Farhan et al., 2016b; Morshed et al., 2019; Adler et al., 2020; Saha et al., 2021; Das Swain et al., 2022]. For instance, Saeb et al. [2015] and Wang et al. [2014] showed that features such as location entropy, sleep regularity, and phone usage could be predictive of depressive symptoms. Canzian and Musolesi [2015] used mobility data to infer depressive states, while Farhan et al. [2016] integrated contextual data (e.g., activity type and weather) to enhance detection accuracy. These studies demonstrate the feasibility of identifying mental health markers from passively sensed data, offering a complementary method to traditional survey-based assessments.

Researchers have also examined how behavioral data reflects physical well-being, such as activity levels, sleep quality, and circadian rhythms. A team of researchers proposed frameworks to extract behavioral features from raw sensing data, enabling prediction of health-related outcomes [Doryab et al., 2014, 2018]. Behavioral sensing has also been employed to capture dimensions of social well-being. Phone logs, Bluetooth proximity, and conversation detection have been used to infer social interactions, detect social withdrawal, and assess loneliness (e.g., [Eagle et al., 2009; Harari et al., 2017]). For example, Wang et al. [2018] analyzed call and messaging patterns to assess variations in social behavior among students under academic stress. Other work has explored how social activity patterns differ across contexts such as work-from-home scenarios or during crises like the COVID-19 pandemic (e.g., [Huckins et al., 2020a]).

Gaps and Opportunities. While behavioral sensing has shown considerable promise for understanding and supporting well-being, important gaps remain in how these technologies are developed, evaluated, and applied—particularly in educational contexts. A key challenge is the limited availability of large-scale, well-documented, and publicly accessible datasets. Many existing studies rely on proprietary or short-term datasets, making it difficult to reproduce results, benchmark models, or compare findings across contexts. Publicly available datasets—especially those that span multiple quarters or academic years—are crucial for advancing longitudinal research, training generalizable models, and supporting open innovation in student-centered sensing

systems.

Another critical opportunity lies in better integrating behavioral sensing with qualitative insights. While passive data, combined with self-reports, can capture patterns of activity and engagement, it often lacks the contextual nuance necessary to interpret why certain behaviors occur or how they are experienced by students. Incorporating qualitative data—especially through direct engagement with students—can reveal how stress, academic pressures, or personal identities shape college life. Such mixed-methods approaches enable a deeper understanding of students' dynamic and diverse needs, and can guide the design of more personalized, context-aware, and effective intervention tools that better support students' academic and emotional well-being.

Impact of COVID-19 on Disability & Mental Health in Higher Education

Disability and Mental Health in Higher Education. Accurate statistics on college students with disabilities are difficult to obtain due to underreporting and inconsistent data collection [Evans et al., 2017; Blaser and Ladner, 2020]. Nonetheless, learning disabilities, ADHD, mental health conditions, and chronic illnesses are among the most commonly reported disabilities in higher education [Price, 2011; Cortiella and Horowitz, 2014; Evans et al., 2017; Lipka et al., 2020]. Studies estimate that up to 50% of students have one or more of these conditions [Blanco et al., 2008; Price, 2011; Evans et al., 2017; Lipka et al., 2020], yet few access formal accommodations [Evans et al., 2017]. For instance, only 17% of college students with learning disabilities use school-provided assistance, compared to 94% in high school [Cortiella and Horowitz, 2014]. Similarly, while nearly 30% of students report receiving mental health diagnoses or treatment in the past year [Association et al., 2015; Bravo et al., 2018], many do not seek disability services. Structural barriers, including inaccessible processes and environments, contribute to lower enrollment, delayed entry, and lower completion rates for students with disabilities [Wolanin and Steele, 2004; Singh, 2019]. The COVID-19 pandemic offers a potential natural experiment to examine how reducing these barriers affects access and success.

The Impact of COVID-19 on Disability and Mental Health. People with disabilities faced heightened risks during the pandemic, including disrupted health care, difficulties with

social distancing, and inaccessible information [Campbell et al., 2009]. These conditions increased vulnerability to stress and distress. Students with pre-existing mental health concerns also encountered intensified challenges. Even before COVID-19, mental health issues were prevalent in college settings [Ketchen Lipson et al., 2015], and prior public health crises (e.g., SARS) have been linked to adverse psychological outcomes [Hawryluck et al., 2004; Lee et al., 2007]. Although some studies found resilience or no widespread increases in distress [Folk et al., 2020; Kanter et al., 2020], others reported worsened mental health among college students during the pandemic [Huckins et al., 2020b; Wang et al., 2020; Fruehwirth et al., 2021; Mack et al., 2021]. Students with a history of chronic stress may be especially vulnerable due to elevated allostatic load, which compromises their capacity to cope [McEwen, 1998].

The Impact of the COVID-19 Pandemic on Education Accessibility. The pandemic-induced shift to online learning created both barriers and opportunities for students with disabilities. While online learning had been growing pre-pandemic [Pei and Wu, 2019], accessibility challenges persisted [Russ and Hamidi, 2021]. Legal accommodations require documentation and self-disclosure, which can be daunting [Blaser and Ladner, 2020]. The rapid transition left little time for faculty to ensure accessible content [Cameron-Standerford et al., 2020]. Technologies considered accessible still posed usability challenges [Byers et al., 2021]. Despite these barriers, remote learning offered unexpected benefits for some students, such as reduced need for disability disclosure, increased flexibility, and fewer physical access issues. A holistic understanding of disabled students' experiences—beyond just course content—is needed to assess the long-term accessibility implications of online education [Tamjeed et al., 2021].

2.3 A Six-Year Study on College Student Experience

2.3.1 Study Design

From 2018 to 2023, we conducted a longitudinal study at the UW to better understand undergraduate students' academic well-being in the context of their daily lives. Our study was inspired by Wang et al.'s *StudentLife* project and was reviewed and approved by the Institutional Review Board (IRB). A more comprehensive overview of the study design is available in the work of

Sefidgar et al. [2019]; below, we describe the components most relevant to this dissertation.

Participants. We recruited undergraduate students through emails, flyers, and social media posts. Eligible participants were full-time students aged 18 or older, who owned an iOS or Android smartphone and were available for the full duration of the study. In the first year of data collection, all participants were first-year college students. In subsequent years, we aimed to recruit more first-year college students, while returning participants from previous cohorts were also invited. Data collection took place during the Spring term (10 weeks) each year, allowing us to control for seasonal variation. Participants could receive up to $245 annually in compensation based on their level of compliance. To capture a broad range of student experiences, we intentionally recruited individuals from diverse backgrounds, including first-generation college students, students with disabilities, and students from underrepresented minority groups.

Procedure. Following an initial screening questionnaire, participants attended an information session where they reviewed the consent form and compensation structure, installed passive sensing software on their smartphones, received a Fitbit Flex 2 for physiological sensing, and completed a demographic survey. Participants were asked to complete two hour-long surveys across the academic year: one at the beginning of Spring term (*pre*) and one following Spring finals (*post*). In 2018, an additional hour-long survey was administered at the end of previous Winter term; this survey was discontinued in subsequent cohorts to reduce participant burden. These surveys collected information on life experiences, coping and self-regulation strategies, health behaviors, and personality traits.

Throughout the Spring term, participants completed EMAs twice weekly. These surveys assessed affect, stress, and experiences of unfair treatment, with each prompt available for approximately 10 hours. To collect more granular data, we administered four EMA prompts per day during two designated weeks each term (weeks 3 and 8). These periods included both typical weeks and weeks leading up to final exams to capture heightened academic stress. To balance participant burden with data richness, we adopted a hybrid sampling strategy, using twice-weekly EMAs in most weeks and increasing frequency only during targeted intensive sampling windows. This approach allowed us to capture behavioral patterns across both weekdays and weekends.

All surveys were administered through Qualtrics [XM, 2002]. We monitored participants' compliance with EMAs and sensor data collection throughout the study, providing follow-up support as needed to troubleshoot technical issues and ensure continuity of data.

2.3.2 Data Collection

Survey Data. Our *pre/post* surveys include a number of questionnaires to cover various aspects of student life, including 1) personality (BFI-10, The Big-Five Inventory-10 [Rammstedt and John, 2007]), 2) physical health (CHIPS, Cohen-Hoberman Inventory of Physical Symptoms [Cohen and Hoberman, 1983]), 3) mental well-being (e.g., BDI-II, Beck Depression Inventory-II [Beck et al., 1996a]; ERQ, Emotion Regulation Questionnaire [Gross and John, 2003]; UCLA, Short-form UCLA Loneliness Scale [Russell, 1996], BRS, Brief Reslilience Scale [Smith et al., 2008]; RRQ, Rumination-Reflection Questionnaire [Trapnell and Campbell, 1999]; Brief-COPE, Brief Coping Orientation to Problems Experienced Inventory [Carver, 1997]; FSPWB, Flourishing Scale Psychological Well-Being Scale [Diener et al., 2010]; AE, Efficacy for Self Regulated Learning and Academic Self Efficacy), and 4) social well-being (e.g., Sense of Social and Academic Fit Scale [Walton and Cohen, 2007]; GQ, Gratitude Questionnaire [McCullough et al., 2002]; EDS, Everyday Discrimination Scale [Williams et al., 1997]; CEDH, Chronic Work Discrimination and Harassment [Williams et al., 1997; Bobo et al., 2000]).

Our EMAs focus on capturing participants' recent sense of their mental health, including PHQ-4, Patient Health Questionnaire 4 [Kroenke et al., 2009]; PSS-4, Perceived Stress Scale 4 [Cohen et al., 1983]; and PANAS, Positive and Negative Affect Schedule [Watson et al., 1988]. Details of each questionnaire are provided in TableA.1.

Passive Behavioral Data. We developed a mobile app using the AWARE Framework [Ferreira et al., 2015] that continuously collects location, phone usage (screen status), Bluetooth scans, and call logs. Participants installed the app on smartphones and left it running in the background. In addition, we provided Fitbits to collect their physical activities and sleep behaviors. The mobile app and wearable passively collected sensor data 24×7 during the study. Key features from the mobile sensing data include physical activity states (e.g., stationary, walking,

running), application usage (foreground apps and push notifications), battery status (charging/discharging, battery levels), Bluetooth scans (nearby Bluetooth-enabled devices), call logs (incoming, outgoing, and missed calls), GPS location data, screen status (on/off/lock/unlock), and WiFi interactions (connected and surrounding access points). All these data streams are gathered from both iOS and Android devices to address potential socio-economic bias, as studies suggest that Android users generally have lower socio-economic status compared to Apple users [Jamalova and Constantinovits, 2019].

2.3.3 Data Pre-processing & Feature Engineering

Survey Data Preparation. In contrast to the sensor data, which maintained a consistent set of features across all years, the self-reported and EMA data varied due to the addition, modification, or removal of specific questionnaires over time. To address this, we developed a codebase to automatically pre-process the survey data—handling missing fields, aligning questionnaire formats across cohorts, standardizing variable names and structures, and detecting abnormal or inconsistent values through automated checks. The codebase also supports the calculation of aggregated scores for each self-report scale, following the official scoring instructions provided by the respective instruments. After pre-processing, we retained both individual questionnaire items and their aggregated scores as features for subsequent analysis.

Passive Behavioral Data Preparation. This dissertation employs two feature extraction frameworks for processing passive behavioral data. The first utilizes RAPIDS [Rapids, V.1.6; Vega et al., 2021], an open-source platform that provides a Reproducible Analysis Pipeline for Data Streams. RAPIDS supports feature extraction from data collected via various mobile and wearable devices across multiple time windows. The features extracted using this framework are included in our publicly released datasets.

For analyses focused specifically on student academic performance prediction, we adopted the behavioral feature extraction framework proposed by Doryab et al. [2018], which derives general low-level features that capture daily behaviors such as physical activity, phone usage, travel time, screen time, sleep, and step counts. Table A.2 summarizes the extracted low-level behavioral

Table 2.1: Basic participant demographics of four years of datasets. Race-Based URM stands for under-represented minority (African-American, Latinx, Native American, and Pacific Islander); FirstGen stands for First Generation College students .

Demographics	DS1 (2018)	DS2 (2019)	DS3 (2020)	DS4 (2021)
Asian (%)	82 (52.9%)	102 (46.8%)	74 (54.0%)	104 (53.3%)
Black (%)	5 (3.2%)	6 (2.8%)	3 (2.2%)	4 (2.1%)
White (%)	50 (32.3%)	70 (32.1%)	40 (29.2%)	48 (24.6%)
Race-Based URM (%)	18 (11.6%)	37 (17.0%)	18 (13.1%)	36 (18.5%)
Non-Male (%)	107 (69.0%)	111 (50.9%)	76 (55.5%)	128 (65.6%)
Immigrant (%)	34 (21.9%)	54 (24.8%)	35 (25.5%)	48 (24.6%)
First-Gen College Student (%)	53 (34.2%)	75 (34.4%)	52 (38.0%)	89 (45.6%)
Self-Identified Disabled (%)	–	21 (9.6%)	22 (16.1%)	16 (8.2%)
Total	**155**	**218**	**137**	**195**

features, and additional implementation details are provided in Appendix A.1.1.

For both frameworks, we segmented each day into five time epochs—morning (6am–12pm), afternoon (12pm–6pm), evening (6pm–12am), night (12am–6am), and the full day (24 hours)—to enable finer-grained analysis of student behaviors. Features were computed separately for each epoch and for the full day, with sleep-related features extracted only on a daily basis. For each behavioral metric, we also computed descriptive statistics including maximum, minimum, and standard deviation.

2.3.4 Publicly Available Dataset & Behavioral Change Due to COVID-19

We released the first four years (2018 to 2021) of our dataset, referred to as DS1 through DS4. After removing participants with significant missing data, these four datasets include 155, 218, 137, and 195 participants respectively, totaling 705 person-years and representing 497 unique individuals. The dataset offers strong demographic diversity, with a high representation of non-males, including females and non-binaries (59.9%), immigrants (24.3%), first-generation college students (38.2%), and individuals with disabilities (9.1%). It also reflects broad racial diversity, with the largest groups identifying as Asian (51.3%) and White (29.5%), followed by under-represented minority, including African-American, Latinx, Native American, and Pacific Islander

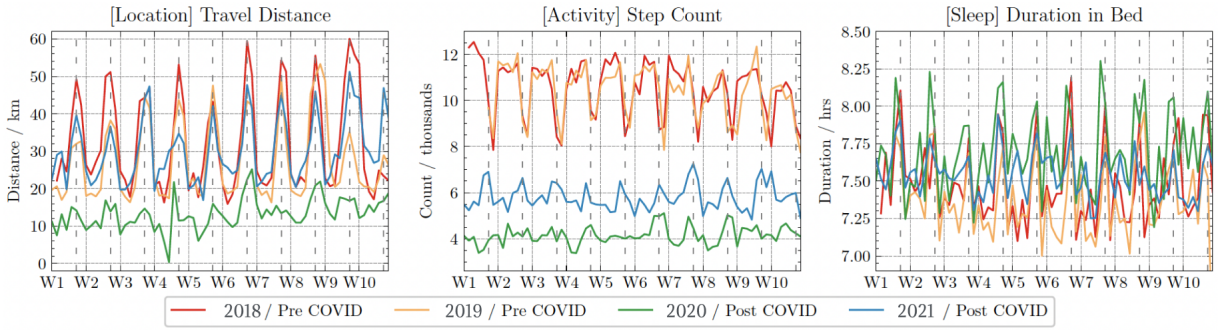


Figure 2.1: Example of behavioral data before and after COVID-19. Grids split weeks. Dashed lines split weekdays/weekends.

(15.6%). Table 2.1 provides a detailed demographic information.

Behavioral Change Due to COVID-19. Because the DS3 data collection period coincided with the initial phase of the national COVID-19 lockdown (March to June 2020) in the U.S., pandemic-related behavioral shifts are clearly reflected in the comparison between DS1&2 and DS3&4, particularly in mobility-related features. For instance, the average daily step count in DS3&4 declined by nearly half relative to pre-pandemic cohorts. However, a recovery trend is observable in DS3&4, as indicated by increases in both travel distance and step counts.

Interestingly, while travel distances in DS4 nearly return to pre-pandemic levels (DS1&2), step counts remain substantially lower. This suggests that participants may have resumed commuting but relied more heavily on non-walking forms of transportation. In addition, weekly behavioral patterns remain evident across all four years. Daily travel distance consistently increases on weekends (particularly Saturdays), while step counts decline—indicating greater movement but less physical activity. Furthermore, participants appear to use weekends to recover sleep, as reflected in increased in-bed duration during that period.

2.4 Student Academic Well-Being and Diverse Needs During COVID-19

In March 2020, during the early phase of our data collection, the first reported COVID-19 death in the U.S. occurred in Seattle—the region where our study was based. Anticipating significant behavioral changes among students (as later evidenced in Figure 2.1), our research team quickly

adapted the survey instruments to better capture the pandemic's impact on students' well-being and academic experiences. Through a series of studies conducted before, during, and after the onset of COVID-19—specifically in March 2020 (immediately following the outbreak), June 2020 (after the transition to online learning), and April 2021 (one year later)—we examined students' mental well-being and their perceptions of online classes and technologies during this period [Morris et al., 2021; Zhang et al., 2022; Nurius et al., 2023]. This section focuses specifically on students with disabilities and mental health concerns, exploring how online learning during the pandemic both challenged and enabled access, using a mixed-methods approach across survey and interview data.

Terminology. In this work, we use the term *students with disabilities/mental health concerns*, which we define broadly to include people who use disability resources and/or self-identify as disabled; and those with mental health concerns who may or may not identify as disabled [Sanderson and Andrews, 2002; Ringland et al., 2019]. We use a modified version of Ringland's inclusion criteria to identify this latter group [Ringland et al., 2019], namely a combination of mental health scales and use of mental health resources. Taken together, we believe this encompasses the most common disabilities reported by college students [Wolanin and Steele, 2004; Price, 2011; Evans et al., 2017], though under-reporting complicates assessment of this [Blanco et al., 2008; Price, 2011].

2.4.1 Methods: Survey & Interviews

To examine the impact of COVID-19 on students, we conducted a quantitative analysis of self-reports, followed by an interview study, leveraging data and participants from the broader longitudinal study described in Section 2.3.

Survey. Our quantitative analysis considers survey data collected in March (*pre-term* survey) and June (*post-term* survey) during COVID-19 in 2020. Students were asked for demographic data including their self-reported disability status. In the *pre-term* and *post-term* surveys, students were asked to complete an hour-long questionnaire consisting of a series of well-established scales to measure chronic discrimination and harassment (CEDH [McGee and Martin, 2011]);

major life events (MLE [Nurius et al., 2021]); perceived stress (PSS [Cohen et al., 1983]); loneliness (UCLA [Russell, 1996]); and perceived social status (SES [Adler et al., 2016]). In addition, we used a post-traumatic stress disorder scale (PCL-5 [Weathers et al., 2013]). Because this scale is only validated as a measure of PTSD when a diagnosis is present, and the questions themselves focus on stress, distress, and their consequences, we refer to this measure as PCL-5 (distress). Two scales (Spring Online Class Concerns/Stress and COVID-19 Related Adversity) listed in Appendix A.2.1 were also included in our questionnaire to examine adversities directly related to COVID-19 pandemic. Several of these scales asked students to reflect over a period of time. The perceived stress and trauma symptomatology assessed the past month. Students reported major life events (adverse, traumatic or stressful events) in the past quarter (Winter quarter in the *pre-term* survey; Spring quarter in the *post-term* survey). Our measures of loneliness did not specify a timescale. It was thus likely that answers to the loneliness scale include current status.

Starting in 2020, a few weeks after the start of the COVID-19 pandemic, courses went online approximately two weeks before the end of the Winter quarter (early March 2020). Students were asked to fill out the *pre-term* survey to assess the immediate reactions to the onset of the COVID-19 pandemic among students with disabilities/mental health concerns compared to their peers, including a range of topics such as pre-existing conditions that could make students vulnerable to COVID-19, and concerns about classes going online. In addition, the study includes a student-specific measure of COVID-19 related adversities. The questions used in the COVID-19 related measures are also shown in Appendix A.2.1.

Spring 2020 was a period of uncertainty marked by the rapid shift to online learning and frequent changes to public health restrictions. At the end of the Spring 2020 (early June), students were asked to fill out the *post-term* survey. Our goal was to understand the differences between students with disabilities/mental health concerns and their peers, before and after a whole quarter of online learning and staying at home. The timeline of data collection and its relationship to COVID-19 is shown in Appendix A.2.2.

Data from 2019 and 2020 were analyzed in [Morris et al., 2021], however no disability specific analysis was completed. The current work explores data from 2020 and specifically compares

students with and without disabilities.

Survey Participants. We assigned students' disability status primarily based on self-identification, while mental health concerns were identified based on a combination of scales from the *pre-term* and *post-term* surveys. Students were categorized as having mental health concerns if they reported *moderate* or *severe* depression[1] (BDI2 [Beck et al., 1996b]) and *severe* anxiety[2] (STAI [Spielberger et al., 1983]) and used mental health resources (e.g., university mental health counseling) within the past year. Our criteria for identifying students with mental health concerns are consistent with classification systems and service provision contexts [Leonardi et al., 2006; Smart, 2011]. Based on our analysis of background literature, such students will likely not identify as disabled, despite facing access issues and experiencing accessibility needs that may overlap with those of self-identified disabled students [Price, 2011].

Our 2020 survey data included 147 students, of which five were excluded from the quantitative analysis due to missing data. Among the 142 students, 81 (57%) were Engineering students, 44 (31%) were Arts & Sciences students, while 17 (12%) students were from other majors. In addition, 65 (46%), 49 (34%), and 28 (20%) students were first-, second- and third-year students, respectively. Of the total sample, 28 were students with disabilities/mental health concerns. Note that students could indicate multiple types of disability and that five students who self-identified as disabled also met the criteria of mental health concerns. Table 2.2 shows the demographics of students in the study.

In March 2020, participants completed the *pre-term* survey after their last Winter quarter final. At that time, the university had been teaching online for at least two weeks, and students knew that half or more of Spring quarter would be online. The majority (80%) of students completed it between March 18 and March 29. At the time, social distancing restrictions were increasing city and state-wide. The initial stay-at-home order in the state where the university is located was issued March 23rd. In June, students were instructed to complete the *post-term* survey after their last Spring quarter final. Nearly all (93%) students completed it between June

[1]In BDI2, a total score of 20-28 is considered moderate depression and 29-63 is severe.

[2]In STAI, a total score of 45-80 is considered severe anxiety.

Table 2.2: Demographics of COVID-19 study survey data. COVID-19 health vulnerability includes students who self-identified as having a pre-existing condition that put them at risk for COVID-19. CUMGPA stands for cumulative GPA of that year. We mark this as N/A for interviews because two interviewees did not provide data on COVID-19 health vulnerability.

Demographics	**Quantitative analysis**		**Interviews**
	D/MH (n=28)	**No D/MH (n=114)**	**D/MH (n=10)**
Female (%)	25 (89.3%)	54 (47.4%)	7 (70%)
Asian (%)	16 (57.1%)	65 (57.0%)	6 (60%)
White (%)	7 (25.0%)	32 (28.0%)	2 (20%)
Black (%)	0 (0%)	2 (1.8%)	0 (0%)
Race-Based URM (%)	3 (10.7%)	12 (10.5%)	2 (20%)
First-Gen (%)	9 (32.1%)	40 (35.1%)	4 (40%)
LGBTQIA+ (%)	8 (28.6%)	15 (13.2%)	2 (20%)
COVID-19 Health Vulnerability (%)	5 (17.9%)	4 (3.5%)	N/A
Average SES	5.5	5.8	4.9
Average CUMGPA	3.43	3.51	3.26
Mental Health Concern	22 (78.6%)		8 (80%)
Vision or Hearing Impairment	10 (35.7%)		1 (10%)
Learning Disability	1 (3.6%)		2 (20%)
Other Disability or Impairment	1 (3.6%)		1 (10%)
Self-Identified Disabled	**19 (67.9%)**		**7 (70%)**

7 and June 14.

Survey Analytic Approach. Data were cleaned and scales were calculated as part of the standard study procedures used in our larger study. Quantitative analysis was guided by the research questions, including full sample portrayals of variable distributions as well as between-group tests of difference (between students with and without disabilities/mental health concerns). When our data were not normally distributed we ran Mann-Whitney U tests (non-parametric) for more conservative significance testing. To address Type-I error rate in multiple comparisons, we used the Benjamini-Hochberg (B-H) method [Benjamini and Hochberg, 1995; Benjamini et al., 2009].

Interviews. Semi-structured individual interviews addressed experiences with remote education and psychosocial well-being during the pandemic. The twenty-seven initial interviews with a general sample, summarized by Morris *et al.* [Morris et al., 2021], were conducted from June 8 to June 24, 2020. An additional four interviews were conducted from March 9 to April 5,

2021. These follow up interviews were exclusively with students with disabilities/mental health concerns and are new to this article.

Interviews were all conducted over Zoom by one or two researchers, and all followed a 90 minute, semi-structured protocol. Participants were asked about a wide variety of aspects of remote education from online classes to study sessions to office hours. Participants discussed technological and social barriers to learning. Participants also were asked about their living situation and how that impacted their education, communication with others and their support networks. Each participant received a $40 Amazon gift card for their participation in the interview. Interviews were recorded and transcribed. An interview guide is included in Appendix A.2.3.

Interview Participants. We draw on two sets of participants for the current analysis. First are six participants who were drawn from the 27 participants interviewed in June of 2020 [Morris et al., 2021]. These six were selected because they met the criteria for disabilities/mental health concerns defined earlier. The data from those students were re-analyzed from an accessibility perspective. To enlarge our sample, we recruited another set of 25 participants who met the same criteria. Four of these 25 students participated in interviews in March/April of 2021. As a supplement to Table 2.2, Table 2.3 details each participant's disability status. To preserve privacy, we do not specify ethnicity or preferred pronouns in the table, but summarize them in the caption. For similar reasons, we use they/them pronouns when referring to participants in the results section. Five interviewees were Asian; two were Latina, and two were white. Three identified as male and seven as female. Participants S2021-MH-31, S2021-MH-32, S2021-ADHD-33 and S2021-VI-34 are the sample from 2021, as specified in their participant ID.

Interview Analytic Approach. As part of thematic analysis conducted for the foundational study [Morris et al., 2021], a preliminary set of codes and related themes were identified. After conducting the four additional interviews in 2021, the researchers developed and applied additional accessibility related codes. In both sets of interviews, notes and interview summaries were written and discussed prior to code development. The researchers resolved all discrepancies in coding through discussion and transcript review. The results presented in this article integrate the disabilities/mental health concerns related codes, including advantages of online learning,

Table 2.3: Demographics of ten interviews. Six interviewees were Asian; two were Latina, and two were white. Three identified as male and seven as female. POTS stands for Postural Orthostatic Tachycardia Syndrome, a complex chronic condition that includes time-varying fatigue and cognitive impacts.

ID #	**Primary D/MH Concern**
S2020-MH-02	Depression
S2020-ADHD-06	ADHD Symptoms
S2020-POTS-09	POTS
S2020-MH-17	Depression
S2020-MH-19	Anxiety
S2020-MH-20	Depression
S2021-MH-31	Anxiety
S2021-MH-32	Depression
S2021-ADHD-33	ADHD
S2021-VI-34	Visual impairment

difficulty/additional burdens of online learning, pandemic exacerbation (examples where pandemic related issues compounded the struggles experienced by students with disabilities) and feeling unworthy of assistance.

2.4.2 Main Results

Survey Results. We first compare students with disabilities/mental health concerns to their peers in the *pre-term* and *post-term* surveys regarding their adversity exposures directly related to the COVID-19 pandemic such as concerns about classes going online. We then examine differences in discrimination, recent major life adverse events, and loneliness.

Online Learning Concerns/Stress and COVID-19 Related Adversities. As shown in Table 2.4, students with disabilities/mental health concerns entered the Spring quarter of 2020 with significantly higher concerns about their online class experience than their peers. At the start of the term, the survey asked about how concerned students were about anticipated stressors; at the end of the end of the term, the survey asked how stressful online learning had been. As shown in Appendix A.2.1, differences in question wording on the two surveys reflects the change from

future predictions to reflection on the past. the mean response to this prospective question in the *pre-term* survey was 2.43 compared to .91 for the retrospective question in the *post-term* survey. A Mann-Whitney test of these *pre-term* and *post-term* responses was significant ($u = 52.0$, $p < .001$) indicating that students' actual experience of online learning (*post-term*) was less challenging than they anticipated for the Spring quarter at *pre-term*.

Table 2.4: Statistics of scores on seven measures of interest for students with disabilities/mental health concerns (D/MH) and without disabilities/mental health concerns (No D/MH) in the *pre-term* and *post-term* surveys. Mann-Whitney U test values and significance levels are indicated. Significance is marked * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. q values are the adjusted p values by B-H method. Measures that are significant after B-H corrections are in bold.

(a) Pre-term survey

Measure	**D/MH**		**No D/MH**		***u***	***p* value**	***q* value**
	Mean	**Std**	**Mean**	**Std**			
Spring Online Classes Concerns	2.43	0.71	2.07	0.76	1146.5	.011*	.015
COVID-19 Related Adversity	2.64	1.77	1.61	1.38	1061.0	.003**	.009
CEDH (Chronic Discrimination)	1.54	1.4	0.87	1.4	1082.5	.002**	.004
Major Life Event (previous quarter)	4.46	3.25	3.13	1.96	1193.5	.018*	.026
PSS (stress)	21.86	6.77	18.32	5.36	1054.0	.003**	.006
PCL-5 (distress)	23.71	14.09	12.63	11.61	810.5	<.001***	<.001
UCLA (Loneliness)	22.18	5.4	22.26	5.05	1588.0	.485	.485

(b) Post-term survey

Measure	**D/MH**		**No D/MH**		***u***	***p* value**	***q* value**
	Mean	**Std**	**Mean**	**Std**			
Spring Online Classes Stress	0.91	0.7	0.95	0.75	1555.5	.418	.418
COVID-19 Related Adversity	2.29	1.76	1.58	1.41	1211.5	.022*	.076
CEDH (Chronic Discrimination)	2.07	2.39	1.29	1.86	1257.5	.032*	.056
Major Life Event (previous quarter)	4.68	3.15	3.08	2.28	1057.0	.003**	.018
PSS (stress)	20.18	7.52	18.11	6.25	1325.5	.093	.130
PCL-5 (distress)	25.48	16.81	18.09	14.65	1141.0	.024*	.056
UCLA (Loneliness)	22.5	5.46	21.94	4.92	1501.5	.315	.367

Stress Exposures. Probing on the finding of increased COVID-19 related adversities and concerns among students with disabilities reported above, we examine a variety of additional factors that might impact students' stress, mental health, and the accessibility of online learning.

In Table 2.4a, students with disabilities/mental health concerns reported higher levels of chronic discrimination (e.g., experiencing demeaning remarks or forms of unfair treatment), re-

cent negative life events (e.g., a serious interpersonal conflict and/or maltreatment that occurred in the past academic ter), greater perceived stress and distress compared to those without disabilities/mental health concerns. Note that both groups of students reported comparable levels of loneliness.

As can be seen in Table 2.4b, after a whole quarter of online learning and staying at home, students with disabilities/mental health concerns still reported significantly more recent negative life events than their peers. The differences between the groups in COVID-19 related adversities, online classes stress, and distress between the two groups diminished. Some differences in Table 2.4b indicate significance via conventional testing, but not all with the more conservative B-H method. Except for a significant decrease in online classes stress, students with disabilities/mental health concerns maintained comparable levels on all other measures before and at the conclusion of the Spring quarter.

Interview Results. Our interview findings provide a contextualized illustration of struggles and benefits of online learning and the pandemic context experienced by students with disabilities. Given heterogeneity of experiences during the pandemic ([Morris et al., 2021]) qualitative analysis provides insight into contexts and impacts that are not evident in comparisons of group averages. Our findings here illustrate that the shift to online learning due to COVID-19 created both challenges and opportunities for students with disabilities/mental health concerns, paralleling findings form our previous finding with a general student sample [Morris et al., 2021]. Most participants reported that some aspects of their online learning were more accessible than their offline equivalents. At the same time, participants described ways in which online learning intensified problems with attention and mood, and added to their fatigue.

Moreover, many participants described financial stressors that constrained their access to treatments, to adequate live/work spaces, and to college itself. As previously examined, factors such as socioeconomic disadvantage and other marginalizing characteristics are often overlapping and can compound stress effects during remote education [Morris et al., 2021]. Here, we highlight needs and opportunities that are specific to students with disabilities/mental health concerns.

We make note of which findings were specific to the ten students with disabilities/mental

health concerns (♿+) and which were found across the larger general sample (🚻) interviewed in both 2020 and 2021.

♿+🚻 *Online Lectures: Benefits and Barriers to Engagement.* There are accessibility benefits of online lectures because they provide great flexibility in how and where students engage with class materials. For example, a student with Postural Orthostatic Tachycardia Syndrome (POTS) (S2020-POTS-09) found it much easier to attend classes remotely. They had fewer symptoms and less exhaustion when they did not have to physically walk around campus.

> *My illness prevents me from standing for a long time or walking for a long time. So for this quarter with things being online. I don't have to walk around campus from place to place. So that helped me control my symptoms and stuff. So learning has not been too bad, actually, for me...* (S2020-POTS-09)

A student with a visual impairment (S2021-VI-34) described challenges and benefits with online learning. They could bring the online content physically closer to their face, so they could read it more easily but they experienced fatigue and difficulty sustaining focus after staring at a screen for long stretches. It was especially hard for them to engage in recorded lectures that contained slides dense with text and code as opposed to more discussion-oriented lectures:

> *Kind of in a mixed bag because, on one hand, the lecture presentations are way closer to me, so I can see the things a lot more clearly, and I can also rewind or like slow down the video and pause... like one of my computer science lectures, ... it's asynchronous and for that there isn't a lot of visual learning to it it's just typing words on a screen so, then there are a lot of times, where I just like I can't focus on this and I pause and then it's good when I take breaks, but then there's times, where I pause and then I like don't come back for hours or days even.... the biggest thing for me is just sometimes forgetting to finish the lectures because it's difficult to focus on.* (S2021-VI-34)

Several students described ways in which online lectures created barriers to engagement, and in turn, intensified mental health and attentional concerns. For example, a student with ADHD (S2021-ADHD-33) talked about the difficulty of focusing without the presence of other students in a lecture hall. They were easily distracted by other screens at their desk and did not have methods for effectively blocking themselves from those distractions. During the pandemic, this student's trouble focusing became more obvious, not just to them but to the two room mates with whom he shared a small apartment where they all studied in the same room for ten hours every day. They nudged the participant to seek ADHD diagnosis and treatment. This participant's exercise and social contact were severely restricted due to fears of exposing themselves and their roommates and family to the virus. Whereas focus and exercise were concerns for the general sample of students [Morris et al., 2021], for students with ADHD the lack of activity, outdoor time and structure appeared to exacerbate difficulty focusing in online lectures and other aspects of online learning [Neudecker et al., 2019].

Some students with depression said that they felt more disengaged with online learning in general, compared to in-person learning, and felt more depressed as a result of that disengagement (e.g., S2020-MH-20, S2021-MH-32). One of these participants, who was facing cuts to their financial aid, job pressures, and a tense home environment, found it much harder to focus when classes were online. They attributed this in part to not being in the same room as the professor, and not seeing their professors move as they spoke, as well as background noise in the small living space they shared with their parents. In addition, they found themselves engaging in other activities during lectures:

> *At first, it was kind of fun to be sitting in my kitchen. listening to my lectures because I could get up and get food whenever I wanted. But then you know, when the lectures lost my interest, it was too easy to walk away to start folding laundry or baking something.* (S2020-MH-20)

Again, this experience is not unique to students with disabilities, and multiple pressures interfered with this particular student's focus. Finally, the slow pace of live lectures prevented

many students from staying focused, including those with ADHD symptoms. One student with ADHD symptoms explained:

> *Sometimes it's hard to sit through like hours of lectures and watch it. It might just be easier to Google the equation and figure out how it works instead of listening through all of it. And so I didn't watch all of my lectures.* (S2020-ADHD-06)

The mixed reactions to online learning illustrate that what is a helpful accommodation for one student can impede learning for students with other disabilities. And even for a given student, an accommodation that addresses some problems may introduce new challenges or tensions that need to be negotiated.

♿+ *Recorded vs. Live Lectures: Negotiating Trade-offs between Accessibility and Accountability.* Recorded lectures, offered either as an alternative to live lectures, or sometimes the sole mode of instruction, provide more flexibility by allowing students to watch lectures at different times and speeds. Prior to the COVID-19 pandemic, lectures were traditionally conducted in person. However, with the outbreak of the pandemic, institutions transited a large number of lectures from in-person to pre-recorded lectures and live video calls. The flexibility afforded by online lectures was described in accessibility terms by students with attentional problems who could not focus on a slow-moving lecture, those with visual impairments who needed to rewind and pause lectures, and those with depression who had trouble attending early morning classes.

A student with major depression appreciated that recorded lectures had become the norm during the pandemic. Recordings of early morning lectures were no longer something they had to request as an accommodation:

> *During COVID it's definitely been more like almost this norm for teachers to record their lectures, which has been really helpful because I didn't have to really advocate, it was already available, zoom recordings are already available.* (S2021-MH-32)

Others appreciated that they could listen to recorded lectures at their convenience, and at a

speed (e.g., 2x) that kept them focused. A student with ADHD symptoms found their live class much more engaging than recorded lectures but still valued the flexibility of recorded lectures:

> *. . . it's difficult to focus. And so in that case, having an online class was more helpful than a live lecture because I'm going to pause lectures and go at my own pace. I don't have to sit in a classroom for like two and a half hours without being able to leave. . . And so because of like the physical freedom. And like there's no restrictions of having to be at a certain place at a certain time for like two hours. For my mental health–I feel like online classes give me more flexibility.* (S2020-ADHD-06)

Similarly, a student with visual impairment (S2021-VI-34) appreciated that recorded lectures allowed them to rewind lectures and to take breaks so that they could rest their eyes. They also could take as much time as they needed to zoom in on a particular slide and, if necessary, take a photo of it with her phone so that they could enlarge it. As described above, however, they often forgot to come back to the lecture after taking breaks.

Although procrastination was a concern raised by students with and without disabilities/-mental health concerns, there were some benefits of this condensed learning (watching lectures *en masse* right before an exam). One participant with ADHD acknowledged their tendency to procrastinate, but also explained that it allowed them to efficiently grasp the relationship between concepts:

> *yeah I'm a huge procrastinator so most of the time I don't watch it and actually watch a week of lectures maybe a few days before the exam and then try to study everything. I think it would be better if I did it continuously, because I know it is better to learn [that way], but I think I do it really efficiently. I try to read everything through. . . look at all [the] pieces of information that's provided and then understanding, I guess kind of comes with looking at [how] everything's connected.* (S2021-ADHD-33)

Similarly, S2020-POTS-09 appreciated the option of recorded lectures but acknowledged that

they also was prone to abandoning a lecture after taking a break. They pushed themselves to watch the lectures live, and in a practice described by many other students, forced themselves to take notes as a way of staying engaged.

> *Yeah, so I think having a recorded lecture is helpful so I can go back and watch it whenever I want. But most of the time, I just tuned in live because I just feel lazy sometimes like to go back and watch things. So I kind of force myself to be on time to live lectures and take notes that way.* (S2020-POTS-09)

In summary, prerecorded lectures meet the accessibility needs of students with disabilities, even though these students (like others) struggled with staying on task while watching them or did not always watch them in a timely manner. It is problematic when a policy denies these powerful needs, simply to paternalistically prevent procrastination.

♿+ *Participating in Class.* Active participation in class is a critical part of learning, but is not equally accessible to all students in traditional classrooms. As described in our our earlier study [Morris et al., 2021], online classes reduced barriers to participation.

From an accessibility perspective, engaging in classes by asking questions is a positive counter force to the disengagement and detachment described by some participants with mental health concerns. This detachment sometimes compounded depressive feelings (e.g., S2020-MH-20). Some students found ways to ask questions online that reduced intense anxiety about negative judgement from peers. One participant described their approach:

> *. . . [when] class ends and you have a specific question, I would go up to the professor myself, but I would never raise my hand in a class [of] 130 people. You could say it's probably fear of rejection. . . Like what if I asked him a dumb question and then also the professor was like laughing at me and, like everyone is like 'Why [are they] so dumb?'* (S2021-MH-31)

In this case, the participant had used the same tactic in-person (i.e., waiting until after the class officially ended to approach the professor and ask questions) to avoid potential embar-

rassment. The scenario they envisioned, of an entire class laughing at them for asking a dumb question, illustrates how intensely anxiety can escalate self consciousness.

Online classes also supported engagement by providing new accessibility solutions. For example, a student with visual impairments (S2021-VI-34) appreciated that they could ask questions and add emoji as a way of expressing herself in a large class. They opted for a feature that allowed them to do so anonymously. They had not asked questions in large classes previously. Expressing oneself through technologies in this way could progress towards fuller participation.

♿+ *Connecting with Other Students.* To support student interaction, many instructors relied on breakout sessions and group assignments. Some students had positive experiences in these settings, whereas others described accessibility concerns that relate to mental health.

In breakout sessions, where students were assigned to discuss a particular topic, it was common for students to turn off their cameras and even their microphones. Many students were discouraged by this lack of participation, something also found in prior interviews with a general student sample [Morris et al., 2021]. The lack of participation sometimes intensified feelings of isolation among students with mental health concerns.

One student with depression (S2021-MH-32) had a particularly negative reaction. They characterized this withdrawal behavior in breakout sessions as a collective statement of "I'm not here." This sense of alienation from peers mirrored the disconnect they felt socially at home. They interacted with few people outside their household, and were not comfortable expressing their feelings and aspects of their identity to their family. The isolation contributed to their depression, which had intensified over the course of the pandemic. To pay for multiple forms of therapy, they started a job that involved considerable risk of COVID-19 exposure. They juggled the anxieties of this increased risk, for themselves and immunocompromised family members, with their treatment. The unsuccessful efforts at connection in breakout sessions and floundering class discussions seemed to reinforce their feelings of isolation from peers and their family.

Group assignments also ran counter to the needs of some participants with mental health concerns (S2020-MH-02, S2021-MH-31). For some students with mental health concerns, these group dynamics added to anxiety or disengagement, making this a less accessible solution. One

student with depression described their frustration with a group project:

> *Our zoom meetings would take sometimes up to three hours just because either we would be waiting on someone to edit something or someone wouldn't join the call and it was just terrible.... It was either you're waiting on somebody to finish or sometimes someone wouldn't even do it at all. So you're stuck with doing double the work.* (S2020-MH-02)

This student's worries about their grades and their academic standing were intensified by this communication breakdown. Like other first generation college students we interviewed, they were constantly torn between helping their family and focusing on their school work. Disorganized calls with peers on this group project ate away at their valuable study time.

As alluded to in the quote above, unequal contributions in group assignments were another source of stress. Such stressors were entwined with living circumstances and disability. One student who struggled with anxiety had to pull the weight of another student for an entire term, and then engage in the emotional labor of explaining the problem to that student (S2021-MH-31). A first generation college student, they felt constantly torn between caregiving and school. Financial stress and concern about the value of their online classes led them to shift to a community college for one term. Their anxiety entwined with the burden compensating for peers, along with financial, caregiving and time pressures.

♿+ⓘ *Office hours, advising and tutoring.* Office hours and meetings with academic advisors were generally easier for students with disabilities and mental health concerns to attend online as was found in a general student sample [Morris et al., 2021]. For students with disabilities/mental health concerns, however, online office hours directly and positively impacted not only their access to the professor, but also health and access to learning materials and peer interaction.

A student with POTS (S2020-POTS-09) found it much less tiring to meet with their advisor online at a scheduled time than to race across campus to the advisor's office, with no guarantee the advisor would even be in. The physical burden of attending office hours in person can be prohibitive in such circumstances.

A student with a visual impairment (S2021-VI-34) started regularly attending office hours and other campus online tutoring for the first time during this period of online learning. They noted that as long as the instructor used screen share (rather than pointing the camera at a physical whiteboard), they were able to read content and follow along. This participant also felt unsafe walking to in-person office hours due to their race and gender. This example extends previous findings that low vision individuals face increased vulnerability to physical threats and violence [Ahmed et al., 2017].

The benefits of participating in office hours online, for both learning and peer interaction, were noted by students with disabilities and mental health concerns. Connection with professors and peers is valuable for supporting students' mental health but has been difficult to fully achieve in the context of online learning [Murphy et al., 2019; Morris et al., 2021]. For some, however, connecting with other students in office hours was easier online. One student with anxiety described how sharing screens felt more comfortable and less intrusive than physically looking over students' shoulders in in-person office hours (S2021-MH-31):

> *In person we would have people lined up for the coding classes,... like 10 of us, and 10 minutes left.... then it just feels like a little competitive ... a little hostile.... Online is really nice because, when someone has a question ... you can share your screen... You don't have to, all huddle over a tiny little 13 inch laptop right, you can all see it.... Oftentimes another person may have the same question to me.... It reinforces the feeling of community I don't feel so alone in you know feeling down ... but then also it's... really helpful, I feel like there's more knowledge to be shared.* (S2021-MH-31)

There were exceptions. For example, one participant with ADHD still felt self-conscious about attending office hours and preferred to resolve questions on their own (S2021-ADHD-33).

♿+ *Shift from Ask and Approve to a Default of Accessibility.* With the shift to online learning in the context of COVID-19, learning models that provide accommodations became the *default*. Such accommodations were previously granted only after students submitted formal requests

through the campus disability resource office.

Several students with mental health concerns (S2021-MH-32, S2021-ADHD-33) were relieved that they no longer had to request exam accommodations. At our university, policies such as open book exams and longer exam times were generally extended to all students during the period of the study. One student with depression (S2021-MH-32) said that the open book tests helped them work against perfectionistic tendencies. As mentioned earlier, they also no longer had to avoid classes that were inaccessible to them due to early morning times, or to ask for accommodations such as taking a test later in the day. Another student, S2021-ADHD-33, appreciated that many of their exams had been canceled.

Relatedly, the flexibility in grading policies and extensions granted during the pandemic were appreciated (e.g., S2020-MH-17). One student with depression appreciated the compassion their professors showed towards students during the pandemic and hoped this continued after in-person learning resumed.

> *You know this is definitely a unique time in all of our lives so I'm hoping that, even though we won't be in a pandemic anymore eventually... that [we will] all still be willing to be flexible and mindful of students' mental health... I think I've had the best experiences where professors are really acknowledging that we're now in destructive environments just based on circumstance... And that we will have pets, or we will be eating, or we will need bathroom break like because that wouldn't usually happen in a traditional classroom setting and I think professors were not like sticklers.* (S2021-MH-32)

However, not all students with disabilities/mental health concerns benefited from such policy changes. For one student with ADHD symptoms, open-book exams were not easier.

> *I think my performance was a little bit lower than usual. I think my grades will be as well... which is surprising because you think that because it's open book (it would be) easier. But that's not necessarily the case.* (S2020-ADHD-06)

Accommodations offered during in-person learning may not always translate to online learning. The student above (S2020-ADHD-06) may have benefited from open book exams when those were taken in a classroom. But at home they were easily distracted by screens, so an invitation to draw on online sources during an exam could backfire.

Reprieve from requesting accommodations related to exams and attendance is potentially significant since making such requests is itself stressful and potentially an aggravating factor for mental health issues. Some students (e.g., S2021-VI-34) held back from requesting accommodations despite obvious health concerns because they felt their problems were not as severe as those experienced by others.

2.4.3 Discussion

This study identified concerns and benefits of online learning for students with disabilities/mental health concerns. In addition to analysis of survey responses during the COVID-19 pandemic, we add qualitative insights about the risks and benefits described by students with disabilities/mental health concerns. The options for recording lectures, asking questions in class via chat, and holding online office hours open doors for participation. Other aspects of online learning were not as successful, and may have had an especially negative effect on students with disabilities/-mental health concerns. Our data provide a perspective on how universities can accommodate these students in meaningful ways. Below we will discuss three areas for improvement, including a broader consideration of accessibility in online learning, more flexible learning approaches and addressing the surrounding stressors on students with disabilities and mental health concerns.

Learning from Flexible Instruction Models. During the COVID-19 pandemic, online learning became the norm by necessity. As universities and instructors plan return to in-person classes, it is natural to ask what, if anything, we should retain from this experiences.

Students with disabilities that affect vision, voice, and mobility may encounter direct accessibility barriers with online learning. Much of traditional accessibility research is focused on reducing the barriers for those groups, i.e. to create equitable access in the face of stable conditions such as adding alt text to image makes a document more accessible to a blind person

(e.g., [Phillips et al., 2012; Kent, 2015]). However an equally pressing problem for students with disabilities that are common in higher education settings, such as chronic conditions or mental health concerns, is that technologies, or required activities, may aggravate their symptoms in some way.

Our findings suggest a need to consider the accessibility of online learning not only in terms of basic materials access but also in terms of its impacts, both positive and negative, on the embodied disability and educational experiences for this broader range of disabled students. For example, online learning can reduce fatigue, a symptom of POTS, by reducing travel needs. Illustrating more problematic factors, it is hard to disentangle online learning from the physical and social isolation experienced by students. This isolation from peers has been detrimental to learning, and may amplify the disconnection felt by students with disabilities/mental health concerns.

Our quantitative results show that when the university first announced that classes would be online, students with disabilities/mental health concerns worried that the shift to online learning would negatively impact grades, academic requirements and admission into their chosen major, but did not have as many worries after experiencing online learning. Our qualitative results explore this dynamic by illustrating some of the positives of online learning, as well as places where additional support might be needed, across a wide spectrum of disabilities. Overall, our results demonstrate that different students bring different needs to the educational experience, and that there is no single best educational model for students.

For example, building on prior findings that some students find it much easier to ask questions in online classes [Morris et al., 2021], some students with disabilities/mental health concerns reported that they were more comfortable asking questions online because they could do so in various ways without drawing attention to themselves. However, here we emphasize that many students, including students with disabilities/mental health concerns, may need more support to take the risk of speaking up in class. Indeed, some students reported that the inability to ask questions in recorded lectures and delayed responses from professors through ancillary asynchronous tools made their learning more difficult. Others reported frustration due to the lack

of interaction with other students in breakout sessions, exercises that are intended to promote student interaction. In addition, many students with disabilities/mental health concerns felt more isolated and this isolation in turn, contributed to their depression and anxiety. In another example, some students benefited from watching pre-recorded lectures at different times and speeds while others found it hard to stay focused on pre-recorded lectures.

Despite students' heterogeneous responses to online education, we believe that it is possible to improve education based on these findings. Whether in person or online, we see a need for flexible approaches to instruction in future education. For many students with disabilities/mental health concerns as well as other students, online instruction directly addresses access problems. Since it is unsustainable for most instructors and learning systems to offer both in-person and online functions, there is a need for innovation on feasible ways to support students. Even simple steps such as posting slides ahead of lectures, recording lectures when possible and giving extra time for assignments appears from earlier pandemic accommodations to make a meaningful difference for students without undermining benefits of in person teaching. In the case of online learning, instructors may also want to mitigate some of the more negative impacts reported in our study by designing for closer peer-to-peer interaction and adding live discussion to accompany recorded lectures. To support this interaction, online learning could be enhanced if instructors schedule live discussions when they upload pre-recorded materials. Such class structure checks could be incorporated into accessibility checking tools in addition to checks for alt text and captions.

Disability in context: Relating life stressors to disabled students' experiences. Similar to [Ringland et al., 2019], our study demonstrates the importance of viewing the whole person. Disability is often entwined with other life factors such as poverty, major life events, and other stressors. In making learning accessible, we need to consider student needs holistically.

Our quantitative analysis shows that students with disabilities/mental health concerns report higher overall levels of COVID-19 related adversity than their peers both Pre-term and Post-term (Table 2.2). They also reported higher level of educational concerns in the pandemic context (e.g., admission to preferred major). In addition, students with disabilities/mental health concerns reported higher exposure to major life adversities, and deeper histories of discrimination than their

peers in the pre-term survey. These results indicate an overall higher level of cumulative stress among students with disabilities/mental health concerns which has been associated among undergraduate with mental health distress [Nurius et al., 2021]. Research within the COVID-19 context urges a renewed focus on college student mental health, with findings akin to ours that fear and worry about their own and loves ones' health and well-being add to more normative stress sources alongside effects of isolation, financial and academic uncertainties among college students [Son et al., 2020]. Community-based research within the pandemic has also evidenced significantly higher levels of pandemic-related stressors among adults with disabilities, with significant association to greater negative effects on their psychological well-being [Ciciurkaite et al., 2021].

Our qualitative results illustrated similar concerns relating to socioeconomic pressures. Students described financial stress, job pressures, and difficulties doing school work in crowded living spaces. Multiple students are first-generation college students, who played caregiving roles at home. All of this impacted students' mental health as well as their academic engagement. In a final example, a student described the confluence of stressors related to her depression: a growing feeling of isolation, anxiety about exposing herself and family to the virus, and at the same time taking on a job that increased her COVID risk so that she could pay for therapy. Analysis of undergraduate students has established that students with multiple marginalizing statuses (e.g., disability, first generation, low income, being an international or immigrant student, having a sexual orientation other than heterosexual) experience "stacked stressors" that significantly account for greater felt stress, depression and anxiety [Nurius et al., 2021].

It is clear from these examples that we cannot address accessibility concerns without also understanding contributing life factors. For example, providing students with more time on assignments may not be as effective without financial aid or other supports that can reduce their work commitments outside of school. While not all of these things can be solved by technology, they do suggest that as we innovate on the technology front, we will be able to better understand the reception of that technology (and its potential) if we are more inclusive in both the people we recruit to help us study it and the questions that we ask about its value for them.

2.4.4 Summary, Limitation & Future Work

In this study, we used a mixed-methods approach to understand how the COVID-19 pandemic affects students with disabilities/mental health concerns, which helped us to identify patterns of interaction between accessibility concerns, online education, and stressors that students experience. Over the course of the study, the raging pandemic led to increases in secondary stressors (e.g., loss of income, loss of social support structures, lack of healthcare access, caregiving burdens [Pfefferbaum and North, 2020; Shultz, 2020]). These primary and secondary stressors may be particularly detrimental for students with disabilities/mental health concerns. This is unlikely to be the last broad-scale disruptive event students experience. Even more importantly, individual lives sometimes include disruptive events. For all these reasons, increased attention to accessibility in, and the accessibility of, online and hybrid learning is critical.

Our findings show that students with disabilities/mental health concerns were more worried than other students about the outcomes of the unanticipated shift to online learning at the start of the term but not at the end of the term. The dynamic nature of these stressors is reflected in our qualitative results, where students with disabilities/mental health concerns reported some ways in which classes going online facilitated access and in other cases created barriers to access. For example, for some students the lack of interaction with others, entangled with academic, family and financial stressors, exacerbated symptoms of depression and anxiety. Based on these findings, we argue that when university seek to provide online access, they need to consider both material access *and* negative impact on symptoms associated with disabilities. Further, future learning environments should support personalized and flexible learning that includes online components.

Our choice to combine people who identify as disabled with people who report themselves as having mental health concerns is driven by the relatively high numbers of people with impairing health concerns who do not identify as disabled [Hale et al., 2020], but may still experience accessibility barriers. There are philosophical and pedagogical questions raised by this choice about who "counts" as disabled, and this is complicated by disability invisibility [Mullins and Preyde, 2013] and under reporting [Evans et al., 2017]. Our view is that this level of mental

health concern rendered these students vulnerable in the COVID-19 context. That said, we recognize that others may use differing definitions.

One limitation of this study is the small sample size. Findings may not generalize beyond students with the particular disabilities and mental health concerns described by our participants. This reflects recruitment challenges—common in human-subject research involving people with disabilities—rather than thematic saturation. A second limitation is that many disabilities and mental health concerns are not represented in this small sample. For example, our study did not include participants with cognitive impairments, neurodiverse students, and students who are deaf or hard of hearing. A third limitation concerns the two different years of data collection; participants with disabilities and mental health concerns were interviewed in 2020 and 2021, but the general sample was interviewed only in 2020. We note the continuity of themes across these samples, but it is possible that some differences result from the year in which participants were interviewed. As noted in the results, some of the concerns expressed by students with disabilities/mental health concerns, such as isolation, disengagement and procrastination, were also expressed by students in a previous study with a general sample [Morris et al., 2021]. Although these findings overlap, such considerations may be particularly consequential for students with disabilities/mental health concerns. The sense of estrangement that comes with online learning intensifies emotional distress and adds barriers to connectedness for students who already feel marginalized in higher education.

Finally, our study only considered a subset of stakeholders at the university. In addition to students, and instructional faculty and staff, other potential stakeholders who could potentially impact accessibility include university administrators and mental health and social well-being staff. For example, given the rising rates of serious mental illness among college students [Storrie et al., 2010], prevention and resilience-fostering supports and programs that specifically address disability needs (e.g., [Stuntzner and Hartley, 2014]) could be of value.

To summarize, the COVID-19 pandemic's impact on learning should be a wake-up call to accessibility researchers to study online learning technologies and their impacts, and higher education in general, from a disability perspective. Although social distancing may fade into

memory, it is likely that online learning will not. The accessibility gains and challenges the COVID-19 pandemic has spurred must not be erased as we return to in-person learning.

Chapter 3

Investigating Harms in Current Behavioral Sensing Practice

3.1 Overview

Despite the rapid advancement and growing promise of behavioral sensing, most existing work has disproportionately emphasized predictive accuracy, with far less attention to whether the resulting models—and the broader research lifecycle—are equitable, transparent, or socially responsible. This imbalance is especially concerning given that behavioral sensing systems are increasingly applied in high-stakes domains and among populations already exposed to structural vulnerabilities. In this chapter, we critically examine how these technologies—when applied to the same underlying data—can introduce harm. We focus on *what* kinds of harms emerge from model outputs and decisions made throughout the sensing pipeline, *why* such issues persist despite rising awareness within the Human-Computer Interaction (HCI) and ML communities, and *how* they might be addressed through more conscientious design and evaluation. By surfacing these often-overlooked risks, we aim to reorient the field toward more responsible, reflexive, and context-aware approaches to behavioral sensing.

The first half of this chapter presents findings from an interview study with 14 behavioral sensing researchers across academic and industry domains. Through these interviews, we explore how fairness is interpreted and operationalized in practice, the sources of bias encountered across different stages of the sensing pipeline, and the institutional, methodological, and community forces that shape ethical decision-making. By grounding our inquiry in practitioners' lived experiences, we move beyond abstract calls for fairness to expose the practical tensions, blind spots,

and trade-offs that researchers must navigate in the development and deployment of behavioral sensing systems.

To complement this insights, the second half of the chapter provides empirical evidence of bias embedded in existing behavioral prediction models. Drawing on the first four years of data from our longitudinal study, we re-implement and adapt a series of behavioral sensing algorithms originally developed for depression detection [Xu et al., 2022b, 2023]. We systematically evaluate these models across demographic subgroups (e.g., gender, race, first-generation college status, and sexual minority identity) and situational factors (e.g., ambient temperature, campus location) [Zhang et al., 2024], revealing disparate performance patterns that may disadvantage already marginalized student populations. These findings underscore the limitations of accuracy-centric evaluation and highlight the necessity of fairness-, generalizability-, and reliability-oriented assessments for real-world deployment.

3.2 Background and Related Work

Limitations of Top-Down Sensing Design. Behavioral sensing technologies hold tremendous potential to support human well-being through scalable and real-time data collection. However, many of these systems are developed through a top-down design paradigm, shaped primarily by what is technologically feasible or measurable, rather than the nuanced, lived realities of target users [Mohr et al., 2017a; Brodie et al., 2018]. This approach often embeds implicit assumptions about users' goals, needs, and behaviors—assumptions that may not hold across diverse populations or contexts. The result is a lack of *context sensitivity*, which can lead to misaligned design choices, misinterpretation of behavioral signals, or disengagement. These risks become particularly salient as behavioral sensing systems move from research prototypes to real-world deployment in domains like education and health.

Empirical Evaluations of Algorithmic Harm. Recent work in Ubiquitous Computing (Ubicomp), ML, and HCI has begun to critically examine the fairness and generalizability of behavioral sensing models. For instance, a comprehensive review by Yfantidou et al. [2023] of sensing studies from 2018–2022 found that only 5% explicitly investigated algorithmic harms, and of

those, the majority focused narrowly on gender or age—typically using only accuracy or error rates as evaluation criteria. However, more recent studies have adopted broader fairness frameworks [Suresh and Guttag, 2021] to identify identity-based harms at various stages of the ML lifecycle [Yfantidou et al., 2023b; Zhang et al., 2023a]. For example, Adler et al. demonstrated that behavioral predictors of depression vary in accuracy and consistency across demographic and socioeconomic groups, raising concerns about differential reliability. Other work has shown how models trained in one region or cultural context often fail to generalize elsewhere, even in tasks like mood inference or activity recognition [Assi et al., 2023; Meegahapola et al., 2023]. These studies reveal both the limitations of current modeling practices and the potential harms they can propagate when deployed without critical scrutiny.

Structural Inequities and Contextual Blind Spots. The lack of representation and contextual understanding in behavioral sensing is not just a technical gap—it reflects deeper structural inequalities. Extensive scholarship in psychology and social work has shown how intersecting axes of identity—such as race, class, disability, and immigration status—shape individuals' lived experiences and access to care or support [Karlsen and Nazroo, 2002; Schmitt and Branscombe, 2002; Frost, 2011; Chou et al., 2012]. Yet these dimensions are often overlooked in behavioral datasets and modeling pipelines. For instance, Hangartner et al. [2021] found that job recruitment platforms systematically reduced visibility for immigrant and minority users, while Blaser et al. [2019] highlighted the absence of disability disclosure in tech company reporting. Autoethnographic accounts such as those by Erete et al. [2021] illustrate how sociotechnical systems often ignore or misrepresent the experiences of multiply marginalized communities [Ellis et al., 2011; Collins, 2019]. These blind spots in behavioral sensing risk embedding systemic biases into seemingly neutral algorithms—especially when developers fail to interrogate whose realities are (or are not) captured.

Gaps in Existing Literature. Although ethical concerns around AI fairness have received increasing attention in the broader ML and HCI communities [Amershi et al., 2019; Commission, 2019; Suresh and Guttag, 2019; Lee et al., 2020; Raji et al., 2020; Acemoglu, 2021; Hutiri and Ding, 2022; Ehsan et al., 2023; Liu et al., 2023; Tahaei et al., 2023; Wang et al., 2023], the behav-

ioral sensing research community has yet to fully engage with these imperatives. To date, only a small number of studies have addressed fairness-related concerns in this domain, and nearly all have focused exclusively on algorithmic fairness—typically assessing disparities in model performance across demographic subgroups. We refer to this class of disparities as *identity-based harms*.

While important, this narrow focus overlooks two critical dimensions of fairness specific to behavioral sensing. First, *situation-based harms*—performance disparities arising from environmental or contextual variation (ee.g., device types, ambient conditions, or infrastructure availability) [Yau and Karim, 2004; Abowd, 2012]—are largely neglected. These harms may not map neatly onto demographic attributes, making them difficult to detect, yet they can meaningfully degrade model generalizability and reliability in real-world deployments. Second, behavioral sensing introduces a broader spectrum of fairness risks that emerge not only at the modeling stage, but across the full sensing lifecycle. Decisions about which behaviors to measure, how to define and interpret them, and how to act upon predictions are all socially and institutionally situated. Each of these steps carries potential for embedding implicit bias or amplifying unintended consequences. Addressing these challenges requires both quantitative evaluations of model performance and qualitative insights into how fairness is shaped by design choices, implementation contexts, and human experiences throughout the pipeline.

3.3 Qualitative Understanding from Behavioral Sensing Researchers

In this section, we conduct an interview study with behavioral sensing researchers to gain a better understanding of fairness concerns throughout the behavioral sensing lifecycle. The findings from this study complement the existing lierature and our empirical analysis described in Section 3.4.

3.3.1 Interview Study Methods

We sought to understand how researchers and practitioners working in behavioral sensing conceptualize, evaluate, and navigate fairness and potential harms in their work. We focused on human stakeholders who actively contribute to the design, implementation, or evaluation of behavioral

sensing systems, spanning both academic and industry settings. Given the limited understanding of fairness from within the behavioral sensing research community itself, we conducted an interview study to surface nuanced, context-specific perspectives that are often difficult to capture through surveys or metric-driven evaluations. We designed the study to elicit both reflective and practice-oriented insights, incorporating activities such as visual mapping of participants' behavioral sensing pipelines, walkthroughs of concrete project examples, think-aloud reflections on fairness guidelines, and feedback on a proposed fairness framework. This approach allowed us to investigate fairness not only as an abstract principle, but as it is understood and operationalized across key stages of the behavioral sensing lifecycle.

Data Collection. We recruited participants using a combination of purposive and network-based sampling. We initially reached out to individuals in our professional networks who had experience conducting behavioral sensing studies across academic and industry settings. To ensure diversity and capture a broad range of perspectives, we intentionally selected participants from multiple countries and with varying levels of domain expertise. We also encouraged participants to share our invitation with relevant colleagues to expand our recruitment pool. Inclusion criteria required participants to be 18 years or older and to have led or contributed to at least one behavioral sensing study.

Interested participants completed a brief pre-survey that collected demographic information and background details related to their involvement in behavioral sensing research. This included their gender, current role, years of experience, and primary application domains. The survey also asked for participants' initial perspectives on fairness-related concerns, which helped inform and tailor our semi-structured interview protocol. In total, we interviewed 14 researchers and practitionersfrom five countries (Australia, U.S., Korea, Canada, and Japan). Each participant received an Amazon gift card as compensation: $40 for interviews lasting 60–85 minutes and $60 for those exceeding 85 minutes. Table 3.1 summarizes key participant demographics and study identifiers. To protect privacy, we report only the continent in which each participant works.

Interview Protocol. Each participant took part in a semi-structured interview conducted remotely via Zoom. All interviews were led by one author and were recorded with participant

Table 3.1: Summary of interviewed participants. The table below provides an overview of the 14 participants included in our study. For each participant, we report self-identified gender, current professional role, country of their workplace, years of experience in behavioral sensing, and primary domain(s) of expertise. Continent of workplace is shown to preserve anonymity.

PID	Gender	Current Role	Continent of Workplace	Years of Experience	Primary Domain(s) of Expertise
PID1	Female	Industry Researcher	Oceania	7+ years	Health & Well-being Monitoring, Privacy, Smart Homes & IoT
PID2	Male	Faculty/PI	Oceania	4-6 years	Health & Well-being Monitoring, Social Behavior & Human Interaction
PID3	Male	Industry Researcher	North America	7+ years	Health & Well-being Monitoring
PID4	Female	PhD Student	Oceania	4-6 years	Health & Well-being Monitoring, Smart Homes & IoT, Social Behavior & Human Interaction
PID5	Male	Postdoctoral Researcher	Asia	4-6 years	Health & Well-being Monitoring, Causal Inference
PID6	Female	PhD Student	Asia	1-3 years	Health & Well-being Monitoring, Social Behavior & Human Interaction
PID7	Male	PhD Student	North America	4-6 years	Health & Well-being Monitoring, Social Behavior & Human Interaction
PID8	Male	Industry Researcher	North America	7+ years	Health & Well-being Monitoring, Workplace Productivity, Smart Homes & IoT, Social Behavior & Human Interaction
PID9	Female	Faculty/PI	Asia	7+ years	Health & Well-being Monitoring, Workplace Productivity, Smart Homes & IoT, Social Behavior & Human Interaction
PID10	Male	PhD Student	North America	4-6 years	Health & Well-being Monitoring, Smart Homes & IoT
PID11	Male	PhD Student	North America	4-6 years	Health & Well-being Monitoring, Social Behavior & Human Interaction
PID12	Male	PhD Student	Oceania	1-3 years	Workplace Productivity, Social Behavior & Human Interaction
PID13	Male	Postdoctoral Researcher	North America	7+ years	Health & Well-being Monitoring, Workplace Productivity, Social Behavior & Human Interaction
PID14	Female	PhD Student	Asia	1-3 years	Health & Well-being Monitoring

consent for transcription purposes. The interview protocol consisted of three main parts.

First, we explored participants' practices as well as experiences across the behavioral sensing pipeline. To ground the discussion, we asked participants to visually map each stage of a behavioral sensing study they had worked on using FigJam. This activity helped prompted context-specific accounts of how fairness and potential harms were—or were not—considered throughout their research. We followed up with questions about participants' own definitions of fairness, perceived sources of bias at each stage of the pipeline, and the barriers that limited their

ability to address fairness during design and evaluation. This portion of the interview focused on uncovering fairness challenges that are specific to behavioral sensing.

Next, we turned to fairness evaluation and mitigation strategies. Participants described any methods, metrics, or processes they had used (or considered using) to assess fairness in their prior work. We asked about concrete instances where fairness-related issues had arisen, how they were identified, and how participants responded. We also invited reflections on the unique challenges introduced by the longitudinal and dynamic nature of behavioral data, and how these shaped participants' approach to fairness.

Finally, we asked participants to reflect on opportunities for improving fairness in future work. Participants revisited and annotated the pipeline diagrams they had created earlier, marking areas where fairness concerns might emerge or where interventions could be most impactful. We concluded by sharing a draft version of our proposed fairness framework shown in Figure B.1 that incorporates fairness considerations into the behavioral sensing lifecycle (e.g., inclusive data collection, fairness-aware algorithms) [Zhang et al., 2023b], and invited open-ended feedback, including critiques, suggestions, and broader reflections on its applicability. The study received approval from our IRB.

Data Analysis. We transcribed all interviews, wrote analytic memos after each session, and documented artifacts that participants shared during or after the interviews. To ensure participant confidentiality, we removed names and other identifying information from all transcripts. Interviews ranged from 70 to 110 minutes (mean = 86, SD = 13); two participants completed their interviews in two sessions due to scheduling constraints. We analyzed the transcripts using Thematic Analysis (TA) [Braun and Clarke, 2006] , following the five-phase process outlined by Braun and Clarke. One author began by reading all transcripts to gain familiarity with the data and drafted an initial codebook grounded in the interview structure and content. Three rounds of collaborative coding followed, during which two authors independently coded the same transcript and resolved discrepancies through a mix of synchronous (Zoom) discussions and asynchronous dialogue. After establishing consistency, one author coded the remaining transcripts in two iterative passes, allowing for refinement and expansion of the codebook as new patterns

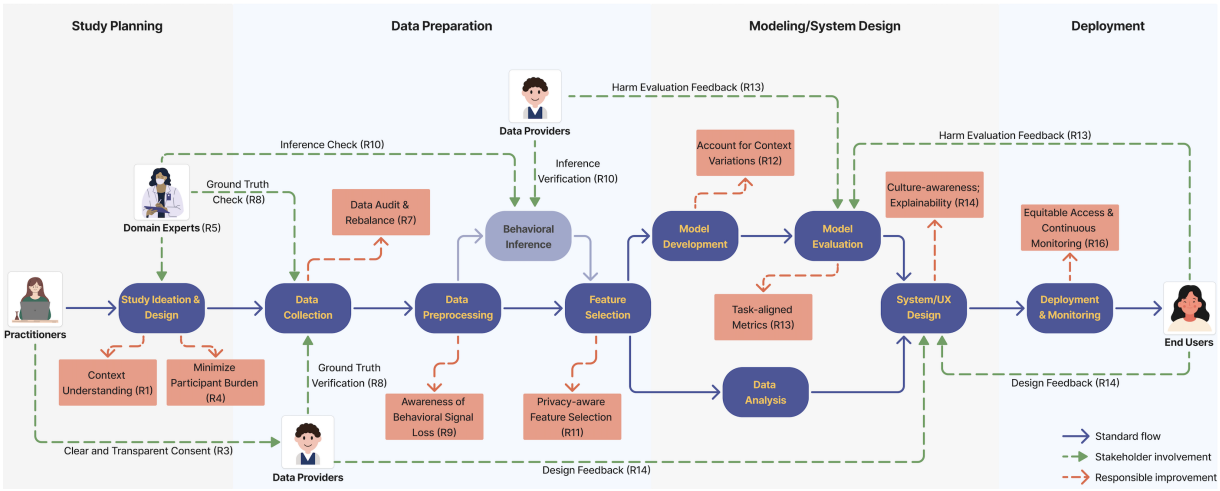


Figure 3.1: A refined workflow for responsible behavioral sensing research, structured into four phases: Study Planning/Data Collection, Data Preparation, Modeling/System Design, and Deployment. Arrows represent standard practices (dark blue), stakeholder feedback needed (green), and responsible fairness-oriented interventions (red). The diagram emphasizes the end-to-end awareness, iterative, and collaborative nature of the process.

emerged. Throughout, we maintained detailed documentation of analytic decisions and codebook iterations to support rigor and transparency.

3.3.2 Findings

We present our findings across three overarching themes, progressing from domain researchers' understandings of fairness in the context of behavioral sensing, to the identification of potential risks and recommended mitigation strategies, and finally to the technical and structural barriers to fair practice. Within each high-level theme, we outline sub-themes that capture distinct yet interrelated dimensions of researchers' experiences and reflections. While the themes are conceptually organized, they are not mutually exclusive; some aspects spread across themes. To avoid confusion, we use the term "researchers" to refer to participants in our interview study or behavioral sensing researchers more broadly, and "participants" to refer to individuals involved in the behavioral sensing studies discussed by those researchers, including both data providers and end users, who may or may not be the same individuals in a given study.

Fairness as Latent, Situated, and Ethical Commitment. Through our interviews, we found that researchers did not consider fairness a necessary part of their prior behavioral sensing studies—at least not when unprompted. However, when fairness was explicitly discussed, their reflections revealed a much broader and more nuanced understanding. Rather than limiting fairness to model performance metrics, researchers emphasized contextual sensitivity, and the need to protect participants' mental and physical well-being.

Fairness Is Rarely a Primary Driver in Behavioral Sensing Practice. When asked to describe a behavioral sensing study they had conducted, most researchers focused on technical and procedural stages—such as study design, data collection, preprocessing, and model development—without mentioning fairness unless directly prompted. Only one researcher (P7), whose work explicitly centered on algorithmic fairness, raised the topic unprompted. As illustrated in Figure 3.1, participants typically followed one of two common pathways (highlighted in dark blue), culminating in either ML–based modeling (11 out of 14) or statistical analysis (3 out of 14), with some extending their work to system design (P5, P6) or deployment (P1, P3, P6, P11). Notably, only two researchers initially mentioned the stage of behavioral inference—a distinct and defining step in behavioral sensing (highlighted in light purple)—yet nearly all (11 out of 14) later acknowledged conducting this stage when fairness-related risks were discussed.

Fairness as Contextual Sensitivity. When we specifically asked researchers about their perceptions of fairness in behavioral sensing, most emphasized that fairness in behavioral sensing cannot be disentangled from the contextual conditions in which sensing practices take place. These conditions include not only participants' demographic attributes such as race (P2, P12), age (P4, P5, P10), gender (P6, P8), health conditions (P9), socioeconomic status (P11). However, researchers stressed that fairness risks also emerge from intersecting structural and situational factors that shape access, participation, and interpretation. Because demographic factors have been widely discussed in both the ML and HCI communities, here we highlight these less frequently examined contextual dimensions.

Access to devices and disparities in technological literacy were among the most frequently cited concerns (P1, P7, P8, P9, P11, P12). Researchers noted that bias can emerge against

individuals who are "*less technologically prone or have less digital literacy*" (P11), leading to skewed data or under-representation. Several researchers also raised concerns about studies that rely exclusively on iOS-based devices, which may systematically exclude individuals who use Android phones (P8, P9, P11, P12). This exclusion, he noted, often intersects with socioeconomic status: "*[It] means that person is probably socioeconomically not that [well]. So you are getting less data from him*" (P11). In addition to concerns about operating systems and socioeconomic exclusion, some researchers pointed out that using uncommon or specialized sensing hardware (e.g., Oura Ring) can further restrict participation. As P7 mentioned, such design choices risk introducing data bias by excluding individuals who fall outside of these narrow user bases.

Beyond device access and literacy, researchers emphasized cultural misalignments as another significant source of fairness breakdowns in behavioral sensing (P2, P4, P6, P9, P14). Several researchers shared examples of model prediction errors that arose from a lack of cultural awareness. P9 described a project in which her team developed models to predict Japanese workers' productivity based on mental health and well-being data. Surprisingly, they found that "*Japanese office workers, even [when] their mental health situation was very bad, their work efficiency was very, very high*," a result that initially confused the team. She later reflected, "*That was very confusing for me why this [was] happening, but then [we realized] this is like a very cultural thing.*" Similarly, P14 shared an example from a study that attempted to infer stress levels among Korean office workers' by using keyboard movements as a proxy for anger. However, the model failed because, as she explained, workers often typed gently even when experiencing stress or anger—behavior she attributed to cultural norms.

Researchers also identified behavior changes—particularly temporal shifts in behavior—as critical contextual factors shaping fairness. Several noted that individuals often modify their behavior in response to being monitored, especially in the early stages of data collection (P2, P8, P10). As P8 explained: "*So what I was getting at was more like when they're being monitored at the start, they might be more conscious of it and try to like purposely alter their behavior to like look better or like be more aware. And then, about two weeks in or so, they start to just forget and then they just actually are being their natural self.*" In addition to short-term behavioral shifts,

researchers also highlighted how the stability of an individual's behavioral routines can influence the kinds of insights that models are able to extract. Participants with more regular routines were perceived as easier to analyze, potentially introducing bias into how models interpret and prioritize certain users. As one researcher explained:

> *"It's like it probably is easier to extract insights for people who have more regularized routines versus not. So we might get biased in sort of the insights. This is the more human element that like we might help a psychotherapist extract. They might be naturally biased towards easier insights to extract which might be based upon people who have more standardized routines."* (P7)

Safeguard for Participants' Mental and Physical Well-being. Many researchers emphasized the importance of protecting participants' mental and physical safety and comfort (P2, P3, P4, P5, P6, P9). Given that behavioral sensing is frequently deployed in high-stakes domains such as emotion regulation and mental health, participants voiced concerns about its potential to cause unintended harm. P2 described emotions as "*so private and intimate*," reflecting on the ethical dilemmas inherent in studying such deeply personal dimensions of human experience: "*I'm always questioning myself if it's good to look into this, because we look deep into human being.*" Others echoed this concern, cautioning that sensing systems—when misaligned with users' lived realities—can heighten distress. As P6 warned, "*If a depressed person sees that, ['Oh, your expression is very depressed,'] then that result can exacerbate the situation. And we can see a worse problem here.*" Beyond emotional risks, some researchers noted that being monitored itself could trigger discomfort. Reflecting on his own study, P5 remarked that participants "*may feel very pressured if they are being watched.*"

In addition to concerns about mental well-being, researchers also highlighted risks to participants' physical well-being—including both discomfort arising from long-term, high-frequency data collection and potential threats to physical safety. For example, P11 reflected on the burden imposed by frequent self-report prompts, such as EMAs used to collect ground truth for sensed data, noting that they can become unsustainable over time: "*Nobody wants to do that.*" These

concerns were further amplified when working with vulnerable populations, such as patients or elder adults, where even minor design oversights could lead to serious consequences. P9 emphasized the importance of minimizing participant burden by carefully determining the number of samples required before initiating a study. She also shared a troubling incident that illustrated how poorly designed sensing hardware could pose safety risks. In one case, an elderly participant with dementia mistook a sensor installed near their bed for food:

> *"We gave them a small installed beacon sensor near their bed to understand their positioning, but one day we discovered that they thought it is a kind of chocolates or something, so it is a kind of very risky [thing], they tried to eat it. So those kind of things are really risky."* (P9)

Potential Risks and Corresponding Mitigation Strategies. When reflecting on potential risks in behavioral sensing, researchers identified a range of potential risks in each stage of the pipeline and proposed corresponding mitigation strategies. Table 3.2 summarizes these key risks and corresponding recommendations, while Figure 3.1 offers a visual overview of our proposed fair behavioral sensing lifecycle. We elaborate on these below, organized into four key phases: study planning, data preparation, modeling and system design, and deployment.

Table 3.2: A lifecycle workbook for risks and mitigations in behavioral sensing. "Depl." refers to the deployment stage. Recommendations are numbered sequentially.

	Stage	Potential Risk	Recommendation (actionable)	
Study Planning	Study Ideation & Design	Limited contextual understanding of target populations and settings.	R1.	Incorporate participatory/context mapping beyond demographics (e.g., monitoring comfort, cultural norms, life events); plan for temporal variability.
		Convenience sampling (university-heavy) may limit external validity.	R2.	Broaden recruitment beyond convenience pools; set inclusion targets and track coverage.

		Data and consent literacy gaps may limit informed autonomy.	R3.	Use plain-language consent with concrete data-use examples; include comprehension checks; integrate fairness/ethics checklists in IRB materials.
		Poor study design can burden participants, harm well-being, and compromise data quality.	R4.	Pilot to minimize burden; consider limits on frequency/duration; randomize/rotate items; include quality checks; schedule with participant context in mind.
		Insufficient domain grounding may lead to misinterpretation.	R5.	Review domain literature and collaborate with domain experts to ensure contextual accuracy and scientific integrity.
Data Preparation	Data Collection	Data logging tool heterogeneity may introduce systematic differences.	R6.	Standardize logging tools/configurations where possible; document versions/capabilities; harmonize pipelines across devices.
		Sampling imbalance may reduce representativeness.	R7.	Audit representativeness iteratively; supplement underrepresented data groups as needed; record recruitment provenance.
		Annotation delays/quality issues may weaken ground truth.	R8.	Use multiple annotators with adjudication; verify labels with participants/domain experts; account for timing offsets.
	Data Pre-processing	Preprocessing choices may introduce bias (e.g., normalization erasing signal).	R9.	Use context-aware preprocessing; Be cautious about pre-processing techniques that can compress meaningful variability and obscure behavioral heterogeneity.
	Behavioral Inference	Unverified behavioral inference may diverge from actual states.	R10.	Triangulate with participant feedback and expert review; encourage researcher-led reflection on inference outputs.
	Feature Selection	Use of privacy-sensitive modalities/features may cause participant discomfort and reduce trust.	R11.	Communicate feature use clearly; offer opt-out/feature-drop options; consider privacy-preserving alternatives when feasible.

Modeling/System Design	Model Development & Model Evaluation	Limited generalizability due to small sample size, cross-population and within-person variations.	R12.	Match model capacity to data scale; evaluate across time and within-subject; consider subgroup models and ensembles; transparently report methodological choices and limitations.
	Model Evaluation	Metric–task misalignment and gaps in fairness evaluation.	R13.	Align metrics with task characteristics; participant-centered evaluation (qualitative and quantitative), develop clearer guidelines for metric selection and evaluation practices.
	Model Development & System Design	Limited explainability and stakeholder usability.	R14.	Use explainable AI; provide stakeholder-tailored, culturally aware explanations with adjustable detail; assess comprehension/engagement; iterate from user feedback.
	System Design	Sensor modality choices may pose safety concerns.	R15.	Select sensors with attention to participants' safety and comfort.
Depl.	Deployment	Representativeness, access, or affordability barriers may reproduce upstream bias.	R16.	Promote equitable access (e.g., low-cost options); monitor performance across groups and adjust accordingly.

Study Planning. In study ideation, researchers noted limited contextual understanding (P1–P5, P7, P8, P10, P11, P14), which can introduce bias early in the research process (P4, P5, P10, P12, P14). Several described reliance on university populations because they are "*easier to approach and recruit*" (P2, P5, P9, P10, P11, P12), a practice that may reduce generalizability—particularly for populations with different lived experiences (e.g., older adults with dementia; P9). Even with demographic diversity, unaccounted factors—participants' comfort with monitoring (P8, P11), cultural norms (P3, P4, P6, P9, P10, P12), and behavioral change over time (P7–P10)—can affect representativeness and interpretation. Researchers also emphasized that psychological and emotional behaviors are "*complex*" and "*more nuanced*," and highlighted situational factors—job status (P2, P11), medical procedures (P3, P5), neurological conditions (P4), and workplace environments (P14)—that shape data. To mitigate these risks, researchers proposed participatory context mapping beyond demographics with explicit planning for temporal variability (R1) and broadening recruitment beyond convenience samples with inclusion targets

and coverage tracking (R2).

Another concern was the gap in data and information literacy during consent, which may limit informed decision-making about data sharing. Researchers suggested using plain-language—explaining what is collected and why through "*some examples*" (P9; R3), clearly stating study goals (P2, P4, P6, P11, P12; R3), and using brief comprehension checks (e.g., "*small tests such as questionnaires*"; P4; R3). They also pointed to transparency around specific data uses (P12; R3), process improvements such as "*developing standardized checklists for fairness and ethical concerns*" (P2, P10; R3), and enhance the clarity and accessibility of consent forms (P2, P9; R3).

Researchers further noted that study burden can affect both data quality and participant well-being (P3, P7, P9). High-frequency or lengthy self-reports may depress compliance and encourage mechanical responding, especially among vulnerable groups such as "*elderly patients*" (P9). To calibrate burden, researchers suggested conducting pre-studies to determine the minimum data needed (P9; R4). They also observed that survey design can degrade data quality over time (e.g., untruthful or rote responses); to mitigate this, P3 recommended "*randomizing the survey questions*" and being "*cautious about study frequency and length*," which can improve response reliability and reduce fatigue (R4).

Given behavioral sensing often intersects with multiple disciplines, researchers stressed the risks of insufficient engagement with relevant disciplinary knowledge. P3 concerned that without incorporating domain expertise, researchers may overlook critical contextual factors or draw flawed conclusions: "*We are the experts of the systems... but the real experts are from the behavioral science domain.*" To ensure appropriate interpretation and ethical grounding, researchers advocated for "*reviewing domain literature*" and "*collaborating with domain experts*" (P1, P3, P6, P7; R5).

Data Preparation. When discussing data collection, researchers revisited earlier concerns about data representative bias introduced by limited contextual understanding (P2, P4, P5, P7, P8, P9, P10, P11, P12). They also noted that variation in data logging tools/software can yield systematic differences: "*Obviously the collection strategies aren't the same—different kinds*

of software, third-party software." To address these issues, researchers recommended auditing dataset coverage iteratively (P1-5, P7-P10, P12, P14; R7) and supplementing underrepresented data where needed (P10, P14; R7). They also suggested standardizing logging tools and software where feasible, documenting their versions and capabilities, and harmonize pipelines across devices (P7, P8; R6).

In cases where participant-reported ground truth was not feasible, some studies relied on human annotators to label behavioral data. However, researchers cautioned that real-world constraints can compromise annotation quality, raising concerns about the reliability of ground truth labels. For instance, P9 described a scenario in which caregivers were asked to manually record the activities of older adult participants. She observed that caregivers often prioritized their caregiving duties—an ethically and practically justified choice—which led to delayed annotations by "*10 minutes or more*." As a consequence, these delays introduced "*a very important issue*" for data accuracy. To address such challenges, researchers recommended verifying labels with participants and domain experts and using multiple annotators to cross-label or validate (P3, P4, P6, P9; R8).

While these strategies aim to improve data collection and labeling, researchers noted that fairness risks can persist during the data preprocessing stage. In particular, they emphasized that routine steps such as normalization or imputation, if applied without consideration of context, may distort or erase meaningful individual differences. For example, P10 cautioned that normalization techniques could compress distinct behavioral patterns: "*You might want to be a bit more careful when normalizing... certain ranges of values get compressed down to a smaller range and then suddenly they all mean the same thing. It can happen based on how you're normalizing*" (R9). Such issues, researchers warned, may lead to biased model interpretations or mask the heterogeneity that behavioral sensing systems are designed to detect.

In contrast to other stages—where researchers often raised fairness concerns unprompted—potential risks associated with behavioral inference were largely revealed through targeted interview questions. When asked about the possibility of mismatches between inferred behaviors and participants' actual states, all researchers who had conducted this stage acknowledged its

challenges. For instance, P4 described a discrepancy between eye-tracking data and self-reported cognitive load:

> "*Two weeks ago, we noticed that one participant's doing very good... and his score is like 63 out of 65, like his score is very good. But he is mentioning that he is in high cognitive load. So in that case, yes, it's not matched with the inferred behaviors.*" (P4)

Despite this awareness, only two researchers reported taking steps to verify or correct such mismatches—typically by incorporating self-reports—before proceeding with modeling or analysis. When asked why verification was not more widely practiced, researchers cited several barriers: a "*lack of domain knowledge*" to interpret discrepancies meaningfully (P1), the limitations of "*quantitative alone couldn't really explain*" the behavior (P8, P11), ambiguity about whether participants' own reports reflected the "*actual situation*" (P11, P12), and the difficulty of intervening in "*real-world settings*" where inference takes place (P14). To address these risks, researchers proposed several strategies beyond quantitative self-reports (P4, P6, P7). These included gathering qualitative feedback through direct conversations with participants and domain experts (P3, P5, P6, P7) and having researchers themselves engage in reflective testing and review of inference outputs (P6, P9).

In the feature selection stage, some researchers raised concerns about the use of sensitive feature modalities such as camera-based sensing (P3), noting that even when such data are collected, participants may still feel uneasy knowing the footage could be further viewed or analyzed. To address these concerns, researchers emphasized the importance of transparency with participants about how each feature would be used (P3, P6). Some also advocated for designing models that allow for the exclusion of certain features—either at participants' request or due to ethical considerations—without significantly degrading model performance. However, they acknowledged that building such flexibility into systems remains a technical challenge (P3).

Beyond privacy, many researchers also questioned the prevailing practice of selecting features primarily based on model accuracy or interpretability (P4–P6, P9–P11). They cautioned that

this approach may lead to the exclusion of contextually important signals from the broader behavioral dataset, thereby limiting the system's capacity to capture the complexity and nuance of human experience. One researcher reflected on this trade-off:

> "*I saw that, like our policies, which is not good. Like we take the feature [based on] feature importance. If we find that these features are not good or these features are actually decreasing our accuracy, we don't consider that these features might be helpful to detect this kind of behavior. But as this feature is decreasing our accuracy, we try to remove that feature... I think there is a potential risk, fairness risk.*" (P4)

Rather than discarding such features outright, researchers suggested carefully examining the underlying reasons for their performance (P4, P9), leveraging explainable AI (P9, P14), and considering whether their exclusion could disproportionately obscure meaningful behavioral signals (P9).

Modeling and System Design. When discussing potential risks in modeling, not surprisingly, researchers unsurprisingly raised concerns about limited generalizability and transparency during model development (P1, P3–P5, P7, P10, P12).In addition to representativeness issues stemming from biased data collection, several researchers highlighted the relatively small sample sizes common in behavioral sensing—especially when compared to large-scale datasets in domains such as Natural Language Processing (NLP) and Computer Vision (CV)—as a further threat to model generalizability (P1, P4, P5, P10, P12). Importantly, researchers emphasized that generalizability should not be considered solely across populations, but also across time at the individual level. As P7 explained in the context of depression prediction:

> "*We said we should focus on... generalization over time, which really starts to get at, like, within-subject fairness. Because people have routine shifts all the time, right? If we're using location entropy to measure depression risk and then the pandemic happens, and everyone's less mobile. That's a distribution shift over time. And... the classifier that worked for them three months ago probably doesn't work for them now.*" (P7)

While these concerns are well recognized in practice, researchers critiqued that they are often underreported in published work (P3). To mitigate these risks, they recommended a range of strategies: conducting multiple validation tests and developing models tailored to specific subgroups that could later be integrated into ensemble approaches (P4, P5); selecting models that balance bias and variance according to dataset size—for example, avoiding "*complex deep learning techniques*" for small samples (P7); and clearly defining the modeling problem by accounting for both individual-level temporal fluctuations and between-person variability (P7). Researchers also emphasized the importance of transparency and honesty in disclosing methodological choices and limitations (P3, P7, P10, P12).

When evaluating models, researchers highlighted concerns about uncritical metric selection during model evaluation and the lack of participant-centered verification (P1, P2, P5–P7, P11). Several noted that standard metrics such as accuracy or F1 score are often used by default, without sufficient reflection on whether these measures adequately capture the nuances of behavioral sensing contexts (P1, P4, P6). Others emphasized that fairness assessments often rely solely on quantitative metrics, overlooking participant perspectives on whether the system's outputs "*feel fair or appropriate*" (P2, P6). To mitigate these risks, researchers advocated for aligning evaluation metrics with the unique characteristics of behavioral data, including contextual sensitivity and class imbalance (P6, P11). However, they also noted a lack of concrete guidance for metric selection and called for "*clearer evaluation guidelines*" (P1, P6, P7). Finally, researchers called for incorporating qualitative feedback, in addition to quantitative evaluation, through more direct engagement with participants during the evaluation process (P2, P6).

For both model development and system design, researchers critiqued the slow adoption of explainability practices in behavioral sensing—despite growing advocacy for transparency and interpretability within the broader HCI and ML/AI communities (P2, P4, P9, P11, P14). Several emphasized that models and systems should be interpretable not only to researchers (P2, P4, P9) but also to key stakeholders, including domain experts (P9) and participants themselves (P4, P9, P14). Researchers further reflected on the socio-cultural dimensions of explanation, highlighting the importance of accounting for "*cultural differences*" in how explanations are perceived (P9,

P10), adjusting the "*information intensity*" to avoid overwhelming participants (P3), and striking a "*balance*" between offering sufficient detail and maintaining participant engagement during explanation delivery (P11). In addition to interpretability, they also called for the "*careful selection of sensors*" to prevent physical harm, particularly for vulnerable populations (P9).

Deployment. When discussing potential risks in deployment, researchers observed that many challenges in this phase mirror those encountered during data collection, particularly the need to ensure representative population coverage (P1, P9, P11). They further emphasized that equitable access to devices and technologies—as well as their affordability—is essential to prevent the exclusion of certain groups from the benefits of behavioral sensing systems (P10, P12).

3.3.3 Summary

Our interview study provide rich, context-specific insights into how fairness and potential harms are perceived, navigated, and—often—overlooked in behavioral sensing research. Domain researchers described concrete examples of risks emerging at different stages of the sensing lifecycle, from problem ideation to deployment, and emphasized the influence of implicit assumptions, data limitations, and structural constraints on fairness-related decisions. We synthesized these insights into a practitioner-ready workbook (Table 3.2) and a visual framework (Figure 3.1) for fair behavioral sensing. While these qualitative accounts offered valuable depth into the why and where of fairness concerns, they could not establish how widespread or systematic such issues might be across existing systems. To bridge this gap, we turned to a quantitative evaluation of existing behavioral sensing pipelines, which we will detail in the next subsection. This approach allowed us to examine whether the concerns voiced by researchers manifest in measurable disparities, to identify patterns of bias that may otherwise go unnoticed, and to ground our reflexive fairness framework in empirical evidence.

3.4 Quantitative Evaluation on Behavioral Models

In this section, we validate the usefulness of our proposed workbook by systematically evaluating nine behavioral sensing pipelines on two tasks: eight depression detection algorithms

re-implemented in our prior work on our publicly available datasets [Xu et al., 2022b], and a regression model for predicting student learning engagement [Gao et al., 2020].

3.4.1 Evaluation Setup

For each evaluation, we start with a background section, delineating the real-world problem, the datasets used, as well as the ML task and algorithms chosen for our evaluation. WE then detail the quantitative fairness evaluation metrics for model performance. This is followed by the evaluation results. Our evaluations mainly focus on two aspects.

- Evaluate the extent to which the potential risks we identified have been considered in previous efforts in the design and implementation of well-being sensing technologies.
- Identify the potential harms and biases these technologies might introduce to users, by performing a quantitative evaluation of those algorithms.

To gain a deeper understanding of algorithmic harms, we further conduct an experiment focusing on bias mitigation in each evaluation study.

3.4.2 Evaluation Study 1: Depression Detection

Background. Research has been conducted using longitudinal passive sensing data from smartphones and wearable devices to predict and detect depression (e.g., [Wahle et al., 2016; Wang et al., 2018a; Xu et al., 2021]). However, these studies often face challenges related to the limited access to datasets and algorithms, hindering reproducibility and transparency in the field. To address these issues, we introduced GLOBEM [Xu et al., 2023], an open-sourced benchmark platform that includes implementations of nine depression detection algorithms and ten domain generalization algorithms. All depression detection algorithms focus on a common binary classification task: distinguishing whether users had at least mild depressive symptoms. In this evaluation study, we examine these depression detection algorithms through a lens focused on potential harms, employing the perspective provided by our proposed framework.

Datasets. We used the complete four-year dataset released from our broader study [Xu et al., 2022b] to evaluate depression detection algorithms. For clarity and comparison, we labeled the

datasets chronologically based on their collection period (D1 to D4). For ground truth labels, we used the BDI-II scale, which was administered once per person at the end of each term.

Depression Detection Algorithms. We evaluated eight depression detection algorithms implemented in our prior work [Xu et al., 2023], which span a range of modeling techniques including support vector machines [Canzian and Musolesi, 2015a; Farhan et al., 2016a; Wahle et al., 2016], logistic regression [Saeb et al., 2015a; Wang et al., 2018a], random forests [Wahle et al., 2016], Adaboost [Xu et al., 2019b], multi-task learning [Lu et al., 2018a], and collaborative filtering [Xu et al., 2021]. One algorithm by Chikersal *et al.* [Chikersal et al., 2021] was excluded due to a substantial discrepancy between our reproduced results and those reported in the implementation work [Xu et al., 2023] (see Table B.1 in the Appendix).

Evaluation Metrics. Below, we detail the evaluation metrics used for study 1.

Criterion 1: Classification Fairness Metrics. We used three fairness metrics: disparity in accuracy, disparity in false negative rate, and disparity in false positive rate. These metrics were applied to assess algorithm performance across individuals with sensitive attributes and those without sensitive attributes. We intentionally chose not to adopt commonly used fairness metrics such as demographic parity (e.g., [Buet-Golfouse and Utyagulov, 2022]), which aim to ensure equal treatment across different groups. This decision was based on prior research findings indicating that individuals with sensitive attributes are more likely to experience depressive symptoms (e.g., [Givens et al., 2007; Lucero et al., 2012; McFadden, 2016]). Using demographic parity, which aims for equal rates of predicted depressive symptoms across groups, could conflict with empirical evidence suggesting inherent disparities in depression prevalence. Our dataset analysis confirmed this, showing notably higher depression levels in certain sensitive groups (first-generation college students, immigrants, and non-male students) from 2018 and 2021[1] (see Figure B.3 in Appendix). This highlights the critical need for selecting fairness metrics that reflect real-world disparities.

Criterion 2: Threshold for Quantifying Differences and Biases. We further added a criterion: a threshold quantifying differences in algorithmic performances across various groups. We imple-

[1]We performed a Mann-Whitney U test with the B-H correction for significance testing.

mented this to mitigate the impact of random variations. For this purpose, we chose established statistical tests, specifically opting for a non-parametric approach, considering the non-normal distribution of the chosen datasets. We utilized the Mann-Whitney U test, a widely recognized method for comparing means between two independent samples, irrespective of their distribution [Mann and Whitney, 1947; Wilcoxon, 1992]. We further employed the B-H correction method to manage the Type I error rate associated with multiple comparisons within the same dataset [Benjamini and Hochberg, 1995]. We set a stringent False Discovery Rate (FDR) threshold at 0.05 [Benjamini and Yekutieli, 2005; Glickman et al., 2014], ensuring that the rate of false positives is carefully controlled at 5%.

Evaluation Results. In our review of nine papers related to the design and implementation of eight depression detection algorithms, we observed that none of the prior work discussed potential harms to users, neither of them engaged with users to better understand their needs. All prior work considered identity-based context, i.e., sensitive attributes. However, consistent with previous sensing technology research, most studies only focused on two sensitive attributes: age [Farhan et al., 2016a; Wahle et al., 2016; Lu et al., 2018a] and gender [Farhan et al., 2016a; Lu et al., 2018a; Wang et al., 2018a; Xu et al., 2021, 2023]. A few also considered race [Farhan et al., 2016a; Lu et al., 2018a; Wang et al., 2018a; Xu et al., 2023], but other sensitive attributes were largely overlooked. In terms of non-demographic aspects, while most studies accounted for data collection time, consideration of device types was less common. Importantly, while these studies reported on this context information, many did not disclose the proportion of data pertaining to each, potentially leading to representative issues. More critically, none of the studies established criteria for evaluating potential harms, nor was there evidence of context-sensitive algorithm design or processes for harm evaluation and mitigation, particularly incorporating user feedback during the whole design process. Furthermore, there was a lack of strategies for the regular maintenance and updating of data and algorithms.

To assess the possibility of potential harms arising from a lack of context sensitivity in these depression detection algorithms, we carried out a quantitative analysis. Specifically, we leveraged the evaluation criteria defined above and evaluated the eight depression detection algorithms

on the five demographic attributes, including gender, first-generation college student status, immigration status, race, and sexual orientation.

Firstly, we observed biases in all algorithms towards certain sensitive attributes, i.e., their disparities in accuracy, false negative rates, and false positive rates. Notably, algorithms with higher balanced accuracy [Xu et al., 2019b, 2022c] tended to show fewer biases across these attributes when evaluated with the three fairness metrics. In particular, the algorithm **Xu_interpretable** [Xu et al., 2019b] did not exhibit bias in terms of accuracy and false positive rate disparities.

Another interesting finding was the reduced bias in all algorithms on DS3, the dataset collected at the start of the COVID-19 outbreak in 2020. This suggests that the significant impact of COVID-19 might have overshadowed other sensitive attributes, leading to this pattern of decreased bias. Additionally, we did not see a consistent pattern indicating which algorithms consistently demonstrated fair performance regarding the sensitive attributes.

3.4.3 Evaluation Study 2: Engagement Prediction

Background. In recent years, addressing the growing concerns of poor academic performance and student disinterest has led to a heightened interest in understanding student engagement, emotions, and daily behavior. This shift has coincided with significant advances in sensing technology, paving the way for novel methods to unobtrusively monitor and analyze student behavior and mental well-being in educational settings. A significant milestone in this domain is the introduction of the *En-Gage* dataset by Gao *et al.* [Gao et al., 2022a]. This dataset is distinguished as the largest and most diverse dataset in environmental and affect sensing within the educational field, offering unparalleled insights into student engagement patterns and classroom dynamics through a diverse array of sensing technologies.

Dataset. The En-Gage dataset includes a four-week cross-sectional study involving 23 Year-10 students (15–17 years old, 13 female and 10 male) and 6 teachers (33–62 years old, four female and two male) in a mixed-gender K12 private school. It utilizes wearable sensors to collect physiological data and daily surveys to gather information on the participants' thermal comfort (the comfort level of students regarding the perceived temperature at the time), learning

Table 3.3: Evaluation results of study 1. Results of algorithmic harms through the disparity in accuracy, the disparity in false negative rate, and the disparity in false positive rate (without incorporating demographic data into the training and testing process). The results are adjusted p-values by Benjamini-Hochberg correction after the Mann-Whitney U test. Significance is highlighted in red. Acc, Fnr, and Fpr are the abbreviations of the disparity in accuracy, the disparity in false negative rate, and the disparity in false positive rate.

Algorithms	**Sensitive Attributes**	**DS1 (2018)**			**DS2 (2019)**			**DS3 (2020)**			**DS4 (2021)**		
		Acc	Fnr	Fpr	Acc	Fnr	Fpr	Acc	Fnr	Fpr	Acc	Fnr	Fpr
Wahle *et al.*	First-gen College Student	0.020	0.030	0.030	0.010	0.050	0.010	0.020	0.030	0.040	0.010	0.050	0.020
	Gender	0.030	0.030	0.030	0.020	0.030	0.030	0.050	0.010	0.020	0.050	0.030	0.050
	Immigration Status	0.040	0.030	0.030	0.040	0.040	0.050	0.010	0.020	0.040	0.040	0.010	0.030
	Race	0.010	0.030	0.030	0.030	0.010	0.020	0.030	0.050	0.030	0.020	0.020	0.010
	Sexual Orientation	0.050	0.030	0.030	0.050	0.020	0.040	0.040	0.040	0.010	0.030	0.040	0.040
Saeb *et al.*	First-gen College Student	0.010	0.030	0.010	0.020	0.030	0.030	0.010	0.050	0.050	0.020	0.030	0.050
	Gender	0.050	0.040	0.050	0.030	0.030	0.030	0.020	0.030	0.020	0.040	0.040	0.030
	Immigration Status	0.020	0.010	0.040	0.010	0.030	0.030	0.040	0.040	0.030	0.050	0.050	0.020
	Race	0.030	0.050	0.020	0.040	0.030	0.030	0.050	0.010	0.010	0.030	0.020	0.010
	Sexual Orientation	0.040	0.020	0.030	0.050	0.030	0.030	0.030	0.020	0.040	0.010	0.010	0.040
Farhan *et al.*	First-gen College Student	0.030	0.020	0.040	0.020	0.030	0.040	0.030	0.040	0.040	0.030	0.010	0.010
	Gender	0.020	0.030	0.030	0.010	0.020	0.010	0.040	0.030	0.050	0.010	0.040	0.030
	Immigration Status	0.040	0.040	0.020	0.040	0.010	0.030	0.050	0.010	0.030	0.050	0.050	0.050
	Race	0.050	0.050	0.010	0.050	0.050	0.050	0.010	0.020	0.010	0.040	0.020	0.040
	Sexual Orientation	0.010	0.010	0.050	0.030	0.040	0.020	0.020	0.050	0.020	0.020	0.030	0.020
Canzian *et al.*	First-gen College Student	0.020	0.030	0.030	0.020	0.010	0.020	0.020	0.020	0.030	0.010	0.030	0.020
	Gender	0.030	0.030	0.030	0.010	0.020	0.010	0.020	0.030	0.020	0.030	0.020	0.030
	Immigration Status	0.040	0.030	0.030	0.030	0.050	0.050	0.050	0.040	0.040	0.050	0.040	0.050
	Race	0.010	0.010	0.030	0.050	0.030	0.040	0.010	0.020	0.010	0.040	0.030	0.040
	Sexual Orientation	0.050	0.030	0.030	0.040	0.030	0.040	0.040	0.050	0.050	0.020	0.050	0.010
Wang *et al.*	First-gen College Student	0.020	0.050	0.040	0.020	0.030	0.030	0.040	0.030	0.050	0.020	0.040	0.030
	Gender	0.040	0.040	0.030	0.030	0.030	0.030	0.050	0.040	0.020	0.040	0.020	0.040
	Immigration Status	0.010	0.020	0.010	0.010	0.030	0.030	0.010	0.010	0.030	0.050	0.010	0.050
	Race	0.030	0.010	0.020	0.040	0.030	0.030	0.020	0.020	0.040	0.030	0.030	0.020
	Sexual Orientation	0.050	0.030	0.050	0.050	0.030	0.030	0.030	0.050	0.010	0.010	0.050	0.010
Lu *et al.*	First-gen College Student	0.010	0.020	0.030	0.010	0.010	0.050	0.040	0.040	0.040	0.050	0.040	0.030
	Gender	0.050	0.030	0.020	0.040	0.020	0.020	0.030	0.040	0.050	0.020	0.030	0.020
	Immigration Status	0.040	0.050	0.040	0.050	0.030	0.040	0.050	0.010	0.020	0.010	0.050	0.010
	Race	0.020	0.010	0.010	0.030	0.040	0.030	0.010	0.020	0.030	0.040	0.010	0.050
	Sexual Orientation	0.030	0.040	0.030	0.020	0.050	0.050	0.020	0.030	0.050	0.030	0.020	0.020
Xu_interpretable *et al.*	First-gen College Student	0.040	0.020	0.050	0.010	0.030	0.010	0.010	0.010	0.020	0.030	0.010	0.030
	Gender	0.020	0.040	0.040	0.020	0.010	0.020	0.040	0.030	0.040	0.010	0.050	0.010
	Immigration Status	0.030	0.010	0.020	0.050	0.050	0.030	0.030	0.020	0.030	0.050	0.030	0.040
	Race	0.010	0.030	0.010	0.030	0.020	0.040	0.020	0.040	0.010	0.040	0.020	0.050
	Sexual Orientation	0.050	0.050	0.030	0.040	0.040	0.050	0.050	0.050	0.050	0.020	0.040	0.020
Xu_personalized *et al.*	First-gen College Student	0.030	0.040	0.020	0.030	0.030	0.040	0.030	0.040	0.020	0.020	0.020	0.010
	Gender	0.010	0.050	0.040	0.040	0.050	0.030	0.050	0.010	0.010	0.050	0.030	0.020
	Immigration Status	0.050	0.020	0.050	0.020	0.040	0.020	0.010	0.030	0.020	0.040	0.050	0.030
	Race	0.020	0.010	0.030	0.010	0.020	0.010	0.020	0.020	0.050	0.030	0.040	0.040
	Sexual Orientation	0.040	0.030	0.010	0.050	0.010	0.050	0.040	0.050	0.040	0.010	0.010	0.050

engagement, seating locations, and emotions during school hours. An initial online survey was conducted to obtain participants' background information, including age, gender, general thermal comfort, and class groups. The dataset reflects the students' organization into different groups (Form group, Math group, and Language group), aiding in tracking their classroom locations. To clarify, students are typically enrolled in courses based on their form group division, except for math courses which are determined by their math group division, and language courses which are determined by their language group division.

Throughout the study, the participants were asked to wear *Empatica E4* wristbands [McCarthy et al., 2016] during school time, which capture 3-axis accelerometer readings, electrodermal activity, photoplethysmography (PPG), and skin temperature. They were also asked to complete online surveys three times a day, posted after certain classes. These surveys capture detailed insights into participants' behavioral, emotional, and cognitive engagement, as well as their emotions, thermal comfort and seating locations [Gao et al., 2022b]. In total, the dataset comprises 291 survey responses and 1415.56 hours of physiological data from all participants.

Engagement Prediction Models. We chose the engagement regression model, LightGBM Regressors [Nemeth et al., 2019], developed by Gao *et al.* [Gao et al., 2020]. The regression model is designed to predict student engagement across three dimensions: *emotional*, *cognitive*, and *behavioral* engagement. Emotional engagement evaluates their feelings of belonging and emotional reaction to the educational environment, cognitive engagement assesses their effort to understand complex ideas and skills, and behavioral engagement looks at students' participation in academic and extracurricular activities. The 1 to 5 Likert scale was used for scoring engagement levels, where 1 represents low and 5 high engagement. To predict these multidimensional scores, a variety of features were extracted, including data from wearable devices and weather stations. It is worth noting that, data such as gender, thermal comfort, and class groups, were not used for the engagement prediction.

Evaluation Metrics. Below, we briefly describe the evaluation methods for study 2.

Given the regression-based prediction task, we used the disparity in Mean Squared Error (MSE) as our primary fairness metric to identify biases in model performance. MSE, the average

of squared discrepancies between predicted and actual values, is widely recognized for assessing regression model accuracy [Wang and Bovik, 2009]. Additionally, to discern if biases were systematic or due to random variation, and considering repeated measurements from individuals, we adopted a linear mixed model method [Laird and Ware, 1982]. This approach involved calculating residuals (differences between actual values and predictions) across various engagement prediction tasks. Subsequently, we utilized a linear mixed model, executed in Python, to examine whether these residuals significantly varied among different groups (e.g., gender and thermal comfort). This statistical method is beneficial for its ability to account for both within-group and between-group variations in the data, thereby offering a deeper insight into the biases present in model performance.

Evaluation Results. In line with the results from our first evaluation study, our examination of relevant papers in this study [Gao et al., 2020, 2022a] indicates that researchers did not engage with the users to understand their needs or considered potential harms to users, and only very limited contextual factors were considered. These factors included gender, thermal comfort at the time of data collection, and information about the courses and classrooms that participants were involved in prior to data collection. Additionally, a key observation is that while this contextual data was considered during the data collection phase, it was not actively incorporated into the training and testing phases of their algorithms. Moreover, the researchers did not address the potential harms of their algorithms. They did not establish criteria for evaluating such harms or implement techniques, including student feedback, to mitigate potential biases. Additionally, there was no evidence of strategies for regular maintenance and updates of the data and algorithms.

We carried out a quantitative analysis to assess the potential negative impacts derived from neglecting certain contextual factors. The findings, detailed in Table 3.4, indicate that specific situated contexts – such as thermal comfort, the group division (e.g., language and math groups), and the courses students were engaged in prior to data collection – significantly influence the performance of the prediction algorithm. For example, as illustrated in Table 3.4, the algorithm's ability to accurately assess emotional engagement was statistically different between students who

Table 3.4: Results of linear mixed models analysis. This table displays the results from linear mixed models, focusing on identifying the significance of differences in regression models across diverse contexts within different engagement prediction tasks. Levels of significance are denoted as follows: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. For each contextual factor, one group is designated as the reference (or baseline) category, for example, the Female group in Gender. The "Interpret" represents the average effect for the reference group when all other predictors are held at their reference level (for categorical variables).

Contextual Factors	**Model Variables**	**Emotional Engagement**			**Cognitive Engagement**			**Behavioral Engagement**		
		Coef.	Std. Error	P> |z|	Coef.	Std. Error	P> |z|	Coef.	Std. Error	P> |z|
Gender	Intercept	0.006	0.119	0.962	-0.055	0.121	0.651	-0.033	0.105	0.753
	groups [T.Male]	-0.025	0.179	0.891	0.080	0.182	0.659	-0.038	0.157	0.810
	Group Var	0.121	0.072		0.120	0.070		0.076	0.044	
Thermal Comfort	Intercept	-0.140	0.116	0.225	-0.113	0.118	0.338	-0.136	0.113	0.232
	groups [T.No change]	0.286	0.112	***0.011**	0.181	0.119	0.130	0.215	0.117	0.067
	groups [T.Warmer]	-0.117	0.150	0.436	-0.013	0.160	0.937	-0.190	0.156	0.225
	Group Var	0.112	0.067		0.098	0.054		0.081	0.052	
Language Group	Intercept	0.001	0.094	0.991	0.101	0.114	0.377	0.021	0.102	0.836
	groups [T.Room 41]	0.721	0.244	****0.003**	-0.492	0.299	0.100	0.327	0.264	0.215
	groups [T.Room 43]	-0.094	0.184	0.611	-0.213	0.219	0.331	-0.248	0.198	0.210
	groups [T.Room 68]	-0.351	0.198	0.076	-0.181	0.241	0.452	-0.324	0.214	0.129
	Group Var	0.057	0.045		0.097	0.070		0.068	0.052	
Math Group	Intercept	0.194	0.166	0.242	-0.314	0.155	***0.043**	0.125	0.153	0.414
	groups [T.Room 41]	-0.292	0.215	0.175	0.330	0.201	0.101	-0.335	0.198	0.091
	groups [T.Room 43]	-0.251	0.223	0.261	0.500	0.209	0.017	-0.131	0.205	0.524
	Group Var	0.111	0.070		0.086	0.05		0.082	0.055	
Course	Intercept	-0.285	0.237	0.228	-0.376	0.241	0.119	-0.022	0.237	0.926
	groups [T.English]	0.488	0.241	***0.043**	0.396	0.246	0.107	0.290	0.246	0.239
	groups [T.Health]	0.400	0.353	0.257	0.395	0.360	0.273	-0.086	0.360	0.811
	groups [T.Language]	0.075	0.255	0.769	-0.148	0.260	0.569	-0.400	0.260	0.124
	groups [T.Maths]	0.274	0.240	0.253	0.530	0.245	***0.030**	0.035	0.245	0.885
	groups [T.PE]	0.485	0.356	0.173	0.558	0.363	0.125	0.226	0.363	0.532
	groups [T.Politics]	0.174	0.251	0.489	0.212	0.256	0.407	-0.343	0.257	0.182
	groups [T.Science]	0.240	0.262	0.360	0.634	0.267	***0.018**	-0.085	0.267	0.749
	Group Var	0.128	0.085		0.126	0.073		0.084	0.054	

were comfortable with the room temperature and those who were not (feeling either too cold or too warm). To delve deeper into this observation, we analyzed the mean squared error (MSE) of the regression algorithm across different levels of thermal comfort. As reported in Table 3.5, the algorithm showed a notably lower error rate ($MSE = 0.631$, $p = 0.011$) when predicting the emotional engagement of students who were comfortable with the temperature, compared to those who were not ($MSE = 0.822$ for students feeling the temperature should be cooler and $MSE = 0.742$ for students feeling the temperature should be warmer). Similarly, our analysis

Table 3.5: Overview of basic statistics. MSE refers to the Mean Squared Error, indicating the average of the squares of the errors. 'Residual' denotes the difference between the ground truth and prediction. MR represents the Mean Residual, which is the average of residuals within each group. "Ind" and "Obs" stand for individuals and observations, respectively.

Context Factors	**Groups**	**Counts (Ind/Obs)**	**Emotional Engagement**		**Cognitive Engagement**		**Behavioral Engagement**	
			MSE	MR	MSE	MR	MSE	MR
Gender	Female	13/149	0.708	-0.014	0.711	-0.005	0.800	-0.023
	Male	10/142	0.693	0.033	0.822	-0.004	0.674	0.002
Thermal Comfort	No Change	22/163	0.631	0.158	0.774	0.069	0.687	0.129
	Cooler	20/77	0.822	-0.140	0.755	-0.135	0.730	-0.115
	Warmer	14/51	0.742	-0.242	0.751	-0.045	0.915	-0.300
Language Group	Room 40	13/155	0.618	0.008	0.690	0.156	0.681	0.051
	Room 41	2/53	0.886	0.703	1.019	-0.563	0.799	0.427
	Room 43	5/52	0.526	-0.075	0.779	-0.082	0.592	-0.186
	Room 68	3/53	1.007	-0.312	0.823	-0.072	1.016	-0.276
Math Group	Room 40	7/80	0.671	0.247	0.849	-0.327	0.640	0.178
	Room 41	9/110	0.763	-0.114	0.715	0.044	0.783	-0.185
	Room 43	7/101	0.657	-0.046	0.753	0.197	0.768	0.030
Course	Chapel	11/12	0.938	-0.268	1.100	-0.297	0.693	0.021
	English	18/71	0.599	0.255	0.484	0.057	0.639	0.340
	Health	8/8	0.991	0.155	1.538	0.149	0.776	-0.035
	Language	20/38	0.986	-0.206	0.970	-0.512	0.975	-0.372
	Maths	20/79	0.551	-0.033	0.816	0.150	0.779	0.0177
	PE	8/8	1.044	0.235	1.314	0.103	0.845	0.224
	Politics	19/43	0.754	-0.119	0.784	-0.101	0.643	-0.341
	Science	19/32	0.641	0.006	0.537	0.252	0.687	-0.050

indicated a significantly higher error rate ($MSE = 0.849$, $p = 0.043$) in predicting the cognitive engagement of students in Room 40 for their math class, as opposed to those in other math groups ($MSE = 0.715$ for Room 41 and $MSE = 0.753$ for Room 43). Interestingly, our analysis revealed no evidence of algorithmic bias or harm, both with gender and in predicting student behavioral engagement.

3.5 Discussion, Limitation & Future Work

In this part, we begin by summarizing the key insights we derived from our mixed-method studies. This summary covers the various findings, their implications, and how they contribute to our understanding of designing behavioral sensing technologies. Following this, we delve into a reflection on our framework, examining its strengths, limitations, and considering perspectives

that extend beyond its current scope.

Potential Harms to Marginalized Groups Due to Context-insensitivity. In both of our evaluation studies, we uncovered a critical and consistent issue with existing behavioral sensing technology designs: a widespread disregard for potential harms to users. Our evaluation revealed that none of the designs thoroughly considered fairness or broader ethical concerns during their design processes. Furthermore, while a few designs did consider the collection of more diverse contextual datasets (e.g., [Xu et al., 2019b, 2021]), this type of data was not utilized effectively during the algorithm training and testing phases. Our quantitative analysis of algorithm performance substantiates the concern of potential harms due to this oversight. Across both studies, we identified performance disparities—whether identity-based or situation-based—that could disproportionately affect certain groups.

For example, our evaluation studies revealed that the emotional engagement prediction algorithm was less effective for students who reported feeling uncomfortably cold or warm, compared to those who were temperature-comfortable (Tables 3.4 and 3.5). While the cause of this disparity remains uncertain, it highlights a potential sensitivity of the model to subjective physical states. Such discomfort may be linked to a range of factors—including chronic illness, medication effects, or variable classroom environments. Regardless of the underlying cause, these findings underscore the need to consider physical and contextual variability—and their uneven distribution—when designing and evaluating behavioral models.

A Need for Engage Users Throughout the Design Process. Another key finding from both of our evaluation studies is the complete absence of user involvement throughout the design process of existing behavioral sensing technologies. Given the widespread use of behavioral sensing technologies, particularly in the mental health domain, this is concerning. As argued by Zhu *et al.* [Zhu et al., 2018], engaging with users in the early stage of the design process can ensure that technologies are designed with a deep understanding of users' needs and values, which can significantly enhance user acceptance and satisfaction. Furthermore, the engagement of users extends beyond the initial design phase to include ongoing feedback loops. Regular interactions with users allow for iterative improvements and adjustments based on evolving needs, emerging

challenges, and changing social contexts [Amershi et al., 2019]. However, it is important to recognize the balance between involving users to mitigate technology harms and minimizing demands on their time and resources. This is especially necessary for people with different needs [Zhang et al., 2022]. Striking this balance ensures that users' contributions are meaningful and sustainable, and that their valuable input genuinely shapes the direction of the technology while respecting their availability and capacity.

Limitations & Future Work. We acknowledge the limited number of participants in our interview study, which may constrain the diversity of perspectives captured and limit the generalizability of our qualitative findings. We also note that our evaluation studies focus exclusively on student populations and well-being-related applications, which may not fully represent the breadth of contexts where behavioral sensing technologies are deployed. Future work should expand to include larger and more demographically varied participant samples, examine additional application domains, and explore whether the patterns of algorithmic bias and ethical risks identified here persist across different settings and populations.

Chapter 4

Building Human-Centered Academic Performance Prediction Models

4.1 Overview

The preceding chapter provide evidence that existing behavioral models carry fairness risks for certain groups. Our follow-up interview study offers insights into how these risks emerge in practice and what design and evaluation strategies can help mitigate them. Building on these lessons, this chapter shifts focus from diagnostic evaluations to forward-looking design: how can we develop predictive models that explicitly incorporate HMCL principles?

In this chapter, we explore modeling strategies that attend not only to accuracy but also to fairness, explainability, and generalizability. We frame the prediction of student academic performance as an early classification task—distinguishing between lower- and higher-performing students—using data available at the beginning of an academic term. We evaluate three modeling approaches designed with societal impact in mind, identify behavioral patterns linked to performance outcomes, and analyze the trade-offs in optimizing for multiple HCML principles.

4.2 Background and Related Work

Need for Early Prediction of At-risk Students in Real-world Settings. Predicting student academic performance has been a longstanding focus of research in educational data mining and learning analytics (e.g., [Daud et al., 2017; Sukhbaatar et al., 2019; Namoun and Alshanqiti, 2020; Khan and Ghosh, 2021; Ojajuni et al., 2021]), and it has recently gained increasing attention in the Computer-Supported Cooperative Work & Social Computing (CSCW)

and broader HCI communities (e.g., [Wang et al., 2014a, 2015a; Sefidgar et al., 2019a; Nepal et al., 2022; Pyle et al., 2023]). A common goal across these studies is to predict end-of-term GPA [Lu et al., 2018b; Sukhbaatar et al., 2019; Chen and Cui, 2020] (as shown in the **Task** column of Table C.1), which, if predicted early enough, can enable timely interventions to improve student outcomes [Cassells, 2018; López Zambrano et al., 2021].

While significant progress has been made, a shared limitation across these efforts is that the earliest predictions occur around a month after the term begins (as indicated in the **Data** column of Table C.1), limiting the opportunity for timely detection and intervention. Timely identification of at-risk students is crucial for providing the best opportunity for support and behavior correction [Hlosta et al., 2017; Cassells, 2018; Thayer et al., 2018]. Research has consistently demonstrated that students who struggle academically early in the term are more likely to experience cumulative negative effects on their performance if these issues are not addressed promptly [Jayaprakash et al., 2014]. Furthermore, the longer it takes to identify students in need of help, the more difficult it becomes to reverse academic difficulties [Katamei and Omwono, 2015]. Early identification allows for the implementation of timely and tailored interventions, such as academic counseling, peer support, and mental health resources, which can help mitigate potential challenges before they escalate [Geiser and Santelices, 2007; Katamei and Omwono, 2015].

In addition to the limitation of delayed early prediction, existing early prediction models also face challenges related to their practical implementation in real-world educational settings. Most prior studies rely on data collected over extended periods, with nearly all requiring end-of-term academic outcomes for model training and evaluation, which inherently limits their ability to be deployed for early interventions. Among all the reviewed work, all but one study, which trained a model on one term's data and applied it to a different term for testing [Chen and Cui, 2020], suffer from this limitation.

Promise of Passive Behavioral Data for Academic Performance Prediction. A recent literature review reveals that most studies in this area rely heavily on data derived from Online Learning Systems (OLS), such as Learning Management Systems (LMS) and Massive

Open Online Courses (MOOC) [López Zambrano et al., 2021]. Our review of prior work on academic performance prediction further supports this finding (as seen in the **Input** column of Table C.1). OLS data typically captures student engagement with course materials, assignment submissions, and exam performance [Mwalumbwe and Mtebe, 2017; Chen and Cui, 2020; Yağcı, 2022]. While this data provides valuable insights into academic engagement, it is often collected over extended periods, which can lead to missed behavioral or performance changes between data collection points [Fredrickson, 2000; Kawakami et al., 2023]. Furthermore, it overlooks crucial daily behaviors and health factors outside the classroom, which can significantly impact academic performance (e.g., [Trockel et al., 2000; Gallagher, 2006; Wyatt and Oswalt, 2013; Felez-Nobrega et al., 2018]).

Daily habits and behaviors such as sleep, physical activity, substance use, and social interactions have been demonstrated to be strongly associated with student academic performance [Trockel et al., 2000; Cox et al., 2007; Wang et al., 2015a; Felez-Nobrega et al., 2018; Xu et al., 2019a]. For example, one study found that weekday and weekend wake-up times had the most significant relative effects on term GPA [Trockel et al., 2000]. Another study demonstrated that time spent in sedentary breaks during weekdays was positively related to academic achievement [Felez-Nobrega et al., 2018]. Additionally, research shows that longer periods of socializing at night, especially as the term progresses, can negatively impact students' term GPA [Wang et al., 2015a]. Beyond daily behaviors, stress and related mental health challenges are growing concerns across colleges and universities, with an increasing number of students experiencing elevated levels of stress and mental health issues [Gallagher, 2006; Hunt and Eisenberg, 2010]. These stressors can significantly undermine academic success, leading to poorer grades, lower GPAs, and higher rates of course withdrawal and dropout [Wyatt and Oswalt, 2013].

Incorporating behavioral and health data offers an opportunity for support systems to proactively identify at-risk students and enable earlier, more effective interventions. With the pervasive and unobtrusive collection of behavioral data, passive sensing data holds the promise of providing continuous insights into students' daily habits and behaviors [Kawakami et al., 2023]. Within the CSCW and broader HCI communities, passive behavioral data has been increasingly studied

and applied in research related to mental well-being (e.g., [Morshed et al., 2019; Sefidgar et al., 2019a; Das Swain et al., 2022; Xu et al., 2023]), as well as social well-being (e.g., [Adler et al., 2022a; Das Swain et al., 2023; Kawakami et al., 2023; Das Swain et al., 2024]). However, the use of such data for predicting student academic performance is still a relatively new area, with only two recognized studies leveraging this approach [Sano et al., 2015; Wang et al., 2015a]. One study focused on predicting end-of-term cumulative GPA using 10 weeks of passive sensing data combined with self-reports [Wang et al., 2015a], while the other classified students in the top 20% and bottom 20% of GPA performers based on one month of similar data [Sano et al., 2015].

Human-centered Nature of Academic Performance Prediction Models. CSCW and HCI researchers have increasingly advocated for integrating technical innovations in ML/AI with human needs and social values [Hong et al., 2020; Kim et al., 2021; Andersen et al., 2023; Banovic et al., 2023; Xu et al., 2023; Zhang et al., 2023a; Meegahapola et al., 2024]. This approach, often referred to as HCML, emphasizes the importance of aligning AI systems with the social and ethical contexts in which they operate. Researchers have pointed out that HCML covers a broad scope, including fair and transparent algorithm design, human-in-the-loop decision-making, human-AI collaboration, and assessing the social impact of ML/AI on diverse communities [Amershi et al., 2019; Chancellor, 2023]. This approach underscores the need for ML/AI systems to be not only technically robust but also sensitive to the broader socio-technical environments they are deployed in.

Compared to research in other domains, such as mental and social well-being, where HCML has become a guiding principle during model and system development [Kim et al., 2021; Andersen et al., 2023; Coleman et al., 2023; Yoo et al., 2024], research in academic support settings has been slower to adopt these principles. While acknowledging the multifaceted nature of HCML, in this review, we focus primarily on examining prior academic performance prediction work in terms of its capability to provide key stakeholders with understandable information and actionable insights for early interventions (*explainability*), ensure equitable deployment across marginalized student groups (*fairness*), and assess the robustness of the models in generalizing across diverse contexts, such as different student populations and academic terms (*generalizability*).

Explainability. While there is still no universal consensus on the exact definition of explainability and related terms such as interpretability [Rosenfeld and Richardson, 2019; Ehsan et al., 2024], explainability, or Explainable AI (XAI), it generally refers to the ability of a system to make its decisions or behaviors understandable to humans [Carvalho et al., 2019; Miller, 2019; Arrieta et al., 2020; Ehsan et al., 2023]. In educational contexts, explainability is particularly critical, as it provides transparency into how predictive models arrive at certain decisions, thereby guiding ML engineers debug the models, and other stakeholders—such as educators, students, and administrators—in working together on potential interventions. By clarifying the model's decision-making process, stakeholders can better understand why certain students are identified as at-risk and what steps can be taken to support them.

One commonly used method for interpreting models is feature importance [Bhatt et al., 2020], which assigns numerical values to each feature based on its influence on the model's predictions. This method helps stakeholders grasp which factors are most significant in determining outcomes, such as students' academic performance (this is the approach we use in this paper). More in-depth descriptions of other explainable or interpretable machine learning methods can be found in works like [Doshi-Velez and Kim, 2017; Du et al., 2019; Molnar, 2020].

In prior work reviewed in Table C.1, several studies touched upon the *explainability* of their models by discussing the factors that contribute to the prediction of students' academic performance [Wang et al., 2015a; Sano et al., 2015; Lu et al., 2018b; Bravo-Agapito et al., 2021; Waheed et al., 2023]. Most identified factors were related to in-class behaviors, including students' interactions with OLS (e.g., visits to forums, attempts at questionnaires) [Lu et al., 2018b; Bravo-Agapito et al., 2021; Waheed et al., 2023], as well as total hours of academic activities [Sano et al., 2015; Wang et al., 2015a]. Additionally, personality traits (e.g., conscientiousness, diligence, and orderliness) have also been identified as factors associated with students' academic performance [Wang et al., 2015a]. While these insights are valuable for understanding the contributors to student performance, they often fall short of translating into actionable interventions. For example, knowing that quiz attempts or forum visits correlate with academic performance does not clarify *how* to intervene. Similarly, personality traits are relatively stable over time

and are not easily influenced through short-term interventions. Without concrete actions tied to these predictors, educators may struggle to design specific interventions that directly address students' needs, limiting the practical utility of these models in real-world educational settings.

Fairness. Exposure to discrimination and social marginalization has long been recognized as contributing to heightened stress levels among students, both directly and through indirect forms such as bearing witness to incidents [Assari et al., 2017; Huynh et al., 2017]. This compounded stress can exacerbate the challenges students face in educational settings, with those experiencing multiple forms of discrimination often showing more pronounced stress responses [Khan et al., 2017]. The persistent strain from stress-related factors disproportionately impacts socially disadvantaged populations, leading to both health and academic performance disparities [Holt et al., 2007; Sternthal et al., 2011; Nadal et al., 2020]. Such disparities further underscore the importance of addressing equity and fairness, especially in systems that impact students' success.

As ML/AI technologies are increasingly integrated into decision-making processes, especially in highly sensitive areas like education, ensuring fairness is crucial. Researchers in HCI, ML, and CSCW have been highlighting the importance of these technologies being *fair*, meaning non-discriminatory with respect to individuals' *protected traits*, such as gender, ethnicity, or religion [Friedler et al., 2016; Chancellor, 2023; Buyl and De Bie, 2024]. Although the field of AI fairness is still developing consensus on both its ontology and methods [Friedler et al., 2016; Verma and Rubin, 2018; Olteanu et al., 2019; Suresh and Guttag, 2019], one well-established source of bias is *algorithmic bias*, where reliance on automated decision-making processes based on ML or statistical methods can amplify biases towards certain subpopulations within the training data [Olteanu et al., 2019].

Algorithmic bias can arise from multiple sources, including model design choices—such as the architecture, loss function, optimizer, and hyper-parameters used [Hooker, 2021; Mehrabi et al., 2021]. These decisions, along with statistically biased estimators, can result in models that are unintentionally introduce or amplify biases, affecting the fairness of outcomes and decision-making processes [Olteanu et al., 2019]. In the context of predictive models in educational settings, such biases may disproportionately affect marginalized student groups, exacerbating

existing inequalities rather than mitigating them.

To quantify these risks, different fairness definitions and metrics have been proposed(e.g., [Xu et al., 2022a]). In this paper, we use three fairness measures that are most applicable to a binary classification setting: demographic parity [Barocas and Selbst, 2016], equalized odds [Zafar et al., 2017], and equal opportunity [Hardt et al., 2016]. A more detailed review of these measures can be found in C.1.1. Despite its significance, fairness in academic performance prediction models remains largely underexplored, with no prior studies we reviewed critically evaluating their models for bias (as shown in the **Model Fairness** column in Table C.1). Our work seeks to address this gap by incorporating fairness evaluation into the development of predictive models.

Generalizability. For real-world deployment in educational settings, ensuring that a model can generalize across different contexts—such as varying populations and institutions—is critical. This means training a model on one dataset and ensuring its accuracy remains robust when tested on one or more previously unseen datasets [Xu et al., 2023]. Generalizability is especially challenging when dealing with longitudinal behavioral data, as behaviors can vary greatly over time (season to season, year to year) and across different locations and individuals. Such variations can alter the data distribution, often leading to a decrease in model accuracy [Adler et al., 2022b; Xu et al., 2023].

Despite the importance of generalizability, almost no academic performance prediction studies address this issue directly. In our review, only one such study—by Chen *et al.* [Chen and Cui, 2020]—evaluates the generalizability of their model. They tested their model on a single class during a new term and found that the AUC dropped from 0.75 to 0.63 on the unseen dataset, demonstrating the challenge of maintaining model performance across different contexts. While there is currently no clear standard defining what constitutes "good" generalizability in terms of AUC scores, a higher AUC on unseen datasets would indicate a model's stronger applicability in varied real-world educational environments.

Summary. To summarize, prior research on academic performance prediction typically focuses on predicting end-of-term GPA using data from OLS. A key limitation of these studies is the delayed identification of at-risk students, as most predictions only occur after collecting data

four weeks into the term. Recent studies from outside the area of academic performance have highlighted the potential of passive behavioral data that could be used to complement OLS data, providing continuous insights that allow for earlier predictions and interventions. In real-world educational settings, the CSCW and HCI communities emphasize the development of systems using a HCML approach, advocating for models that are not only accurate but also explainable, fair, and generalizable. Despite the importance of these considerations, prior work in academic performance prediction has yet to fully address these human-centered challenges. Our work seeks to fill this gap by exploring three approaches to predict, early in the term, whether a student's end-of-term GPA will fall below 3.2, leveraging passive behavioral data collected no later than the first week. We further assess these approaches' explainability, fairness, and generalizability, recognizing that these factors are essential for creating predictive models that align with the human-centered needs of real-world educational environments.

4.3 Data and Ground Truth

We used the data collected no later than the first week of the 2018 and 2019 Spring terms from our larger study to develop our predictive models. We focused on data from these two years to avoid the disruptions to grading and student behavior patterns caused by the COVID-19 pandemic starting in 2020, ensuring consistency and reliability in our modeling approach.

4.3.1 Feature Extraction

In this work, we adopted a different approach from the one described in Section 2.3 to generate the behavioral data used for modeling. We followed the feature extraction framework described in prior work [Doryab et al., 2018] to extract general, low-level behavioral features, including physical activity, phone usage, travel time, screen time, sleep, and step counts. To capture student behaviors with greater granularity, we grouped the daily behavioral data into five time epochs: morning (6am - 12pm), afternoon (12pm - 6pm), evening (6pm - 12am), night (12am - 6am), and an entire day (24 hours), replicating a similar approach employed in prior work [Wang et al., 2015a]. Features were calculated for each epoch, as well as for the entire day. Sleep features

were computed on a daily basis only.

Additionally, we replicated high-level academic-related behavioral features identified in prior research [Wang et al., 2015a; Lu et al., 2010; Lane et al., 2011], such as student activity duration (i.e., non-stationary time), study duration and focus time during study, dorm time, party attendance, indoor and outdoor mobility, class attendance, and changes in behavior patterns. While the calculations of some features were adapted to fit the context and dataset of this study, they remained consistent with the goals of earlier studies. For both the low-level and high-level behavioral features, we computed statistical metrics such as average (avg), standard deviation (std), minimum (min), and maximum (max) values within each epoch. Further details on how each high-level behavioral feature was calculated can be found in Appendix C.1.2.

4.3.2 Self-Reports

In this paper, we used both calculated scale scores and individual items of self-report data as self-report features. The types of passive data collected remains consistent across 2018 and 2019, though the *pre*-term and EMA surveys are slightly different between the two years due to small changes made during the data collection to improve efficiency. A description of a preliminary data cleaning procedures applied to both the behavioral data and self-reports is provided in Appendix C.1.3.

4.3.3 Ground Truth

Students' end-of-term Grade Point Average (GPA) was used as the ground truth for performance classification. The outcome label was binary, distinguishing between students with a GPA above or below a predetermined threshold. Prior research has employed various GPA cutoffs to define at-risk students, such as those predicted to earn a final grade of C+ or lower [Chen and Cui, 2020], those failing a course [Yu et al., 2018; Sukhbaatar et al., 2019], or a GPA below 4.0 on a 6-point scale [Sano et al., 2015]. In this study, a GPA cutoff of 3.2 (on a 4-point scale) was selected, corresponding to both the university's reported average GPA, aligning both with typical

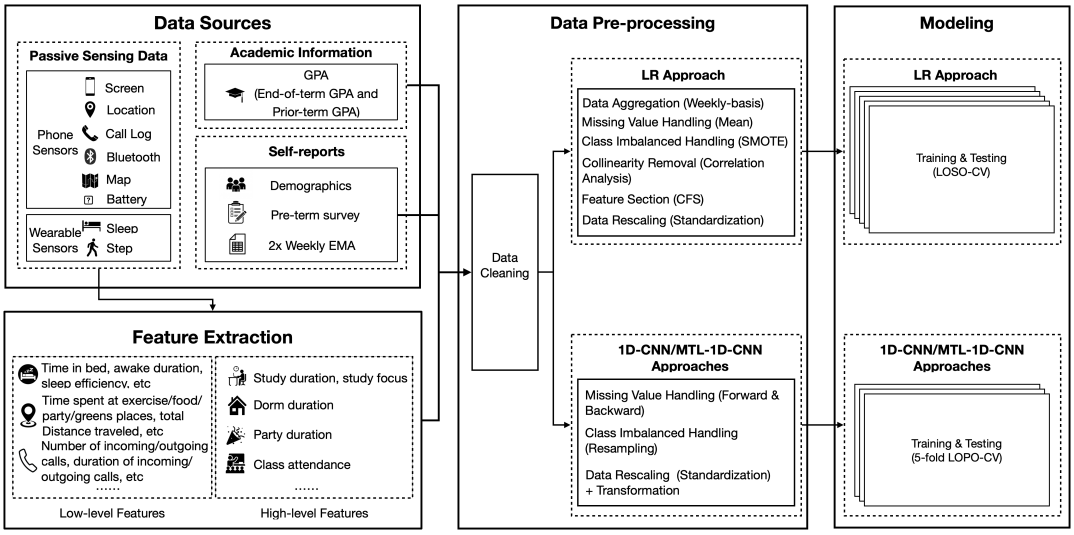


Figure 4.1: Overview of the whole modeling pipeline for the three approaches. All three approaches utilize the same data sources and extracted features. However, distinct data preprocessing and modeling techniques were applied to the LR approach compared to the 1D-CNN and MTL-1D-CNN approaches.

student performance levels and the minimum required for a B+ grade[1]. The left side of Figure 4.1 visualizes all data sources used in this study.

Students with a GPA above 3.2 were classified as high performers, while those with a GPA of 3.2 or lower were considered low performers. In Spring 2018, the mean GPA was 3.48, with a standard deviation of 0.47. Of the 188 students, 145 (77%) were high performers and 43 (23%) were low performers. In Spring 2019, the average GPA decreased slightly to 3.32, with a standard deviation of 0.60. Of the 196 students, 133 (68%) were high performers and 66 (32%) were low performers. The right columns of Table 4.1 offer a breakdown of the number and percentage of low performers across different demographic groups, while the left columns show the total number and percentage of students in both protected and unprotected groups within the overall population. For instance, in 2018, 32 students (17.0%) identified as underrepresented minorities, of whom 12 (37.5%) were categorized as low performers.

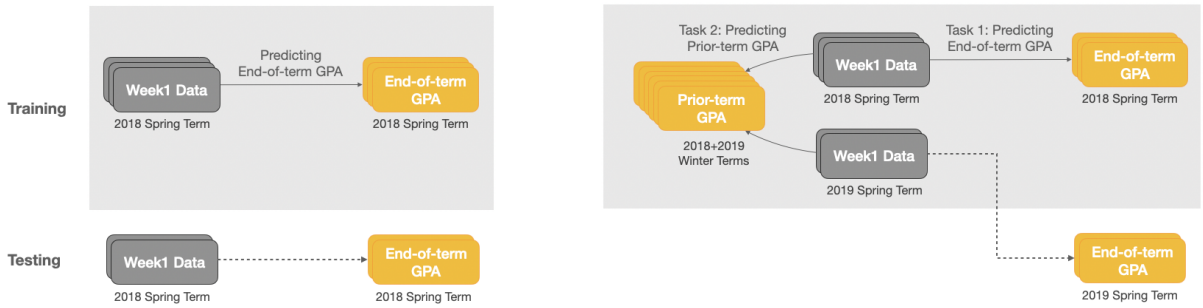
[1]Our data was collected from a smaller subset of students at the university, so the average GPA for each dataset may differ slightly from the overall university average of 3.2.

Table 4.1: Demographics. The table shows the total number and percentage of individuals in the entire population (EP), along with the number of low performers and corresponding percentages within each group: protected group (PG) and unprotected group (UG). Protected groups include underrepresented minorities, first-generation college students, non-male gender (including female and transgender individuals), and sexual minorities (e.g., homosexual and bisexual individuals), based on race, first-generation status, gender, and sexual orientation. "# Total" refers to the total number in each group, and "% in EP" indicates the percentage of that group within the entire population. "# Low Performers" refers to the number of low performers in each group. "% in PG" and "% in UG" represent the percentage of low performers within the protected and unprotected groups, respectively.

Year	**Protected Trait**	**Protected Group (PG)**		**Unprotected Group (UG)**	
		# Total (% in EP)	# Low Performers (% in PG)	# Total (% in EP)	# Low Performers (% in UG)
2018	Race	32 (17.0%)	12 (37.5%)	156 (83.0%)	31 (19.9%)
	First-generation	57 (30.3%)	19 (33.3%)	131 (69.7%)	24 (18.3%)
	Gender	123 (65.4%)	80 (24.4%)	65 (34.5%)	13 (20.0%)
	Sexual Orientation	21 (11.2%)	3 (14.3%)	167 (88.8%)	40 (24.0%)
2019	Race	23 (11.7%)	12 (52.2%)	173 (88.3%)	51 (29.5%)
	First-generation	58 (29.6%)	26 (44.8%)	138 (70.4%)	37 (26.8%)
	Gender	100 (51.0%)	35 (35%)	96 (49.0%)	28 (29.2%)
	Sexual Orientation	21 (10.7%)	5 (23.8%)	175 (89.3%)	58 (33.1%)

4.4 Early Academic Prediction Approaches Within Human-Centered Settings

In this section, we present three early academic prediction modeling approaches, with a focus on the human-centered principles of *explainability*, *fairness*, and *generalizability*. We begin by detailing the design rationale for each approach, including the reasoning behind model selection and the specific methods employed. This is followed by a description of the data pre-processing steps used to prepare the behavioral data, self-reports, and academic outcomes for input into the models. Finally, we provide an overview of the modeling setup, including the specific configurations and methods. The middle and right sections of Figure 4.1 depict the data pre-processing and modeling steps for each of the three modeling approaches. Additionally, Figure 4.2 outlines the training and testing setups specific to these approaches.



(a) LR and 1D-CNN approaches.

(b) MTL-1D-CNN approach.

Figure 4.2: Overview of the training (highlighted in light gray) and testing process for the three approaches. (a) shows the training and testing process for the LR and 1D-CNN approaches (using 2018 Spring term data as an example), where data collected by the first week is used for training and testing to predict end-of-term GPA for both 2018 and 2019. (b) shows the training and testing process for the MTL-1D-CNN approach, where training includes two tasks: Task 1 uses the first week of data from 2018 to predict end-of-term GPA, while Task 2 combines first-week data from both 2018 and 2019 to predict prior-term Winter GPA. Testing uses data from 2019 to predict end-of-term GPA.

4.4.1 The LR Approach

Design Rationale. Logistic Regression (LR) is a widely used and interpretable machine learning model, often selected when explainability is a key priority [Barredo Arrieta et al., 2020]. Its linear structure allows for easy interpretation of feature importance, helping users understand how each variable contributes to the model's predictions [Bhatt et al., 2020]. Additionally, the simplicity of LR makes it easier to address and correct bias, with methods such as reweighing and post-processing bias correction being more effectively applied to linear models, thus supporting fairness across different student groups [Mehrabi et al., 2021].

Data Pre-processing. We implemented several data pre-processing steps to address missing values, class imbalance, collinearity, and feature selection. For missing values, two imputation methods were considered: assigning a default value (999) or imputing based on the mean of the training set, with the latter chosen due to its better performance in the 2018 data experiments. Using the mean value from the training set for both training and testing ensures that test data does not influence the imputation process, preventing data leakage by avoiding any knowledge

of the test distribution [Bishop and Nasrabadi, 2006; Géron, 2022]. To address class imbalance, we employed SMOTE [Chawla et al., 2002], which oversamples the minority class to balance the dataset. Features exhibiting collinearity, indicated by correlations exceeding 0.7 [Dormann et al., 2013], were removed from both the training and test sets to avoid distortion in model estimation. Finally, we applied correlation-based feature selection (CFS) [Hall, 1999], conducting a grid search to identify the optimal correlation cutoff value, r, ensuring generalization from training to unseen data. Additional details on these processes can be found in Appendix C.1.3.

Modeling Setup. We employed Leave-One-Subject-Out Cross-Validation (LOSO-CV) to minimize overfitting and ensure robust model performance. In each iteration of LOSO-CV, feature scaling was applied to the training set, with standardization performed to enhance convergence. All features were aggregated at a weekly level, represented by means and standard deviations, to create a unified data structure.

To fine-tune the details of the LR modeling pipeline, we treated the 2018 data as an "experimental" dataset, testing various methods for missing data imputation, class imbalance handling, different values for the regularization parameter r, and cutoff values for feature selection. Since these experiments could introduce a risk of overfitting to the 2018 data, the final pipeline developed from this experimentation was applied without any modification to the 2019 dataset. Figure 4.2a shows the training and testing process for the LR approach.

4.4.2 The 1D-CNN Approach

Design Rationale. One-Dimensional Convolutional Neural Network (1D-CNN) are a deep learning model that are highly effective for analyzing time-series data, making them particularly advantageous when working with behavioral data collected from sensors [Xu et al., 2023; Kiranyaz et al., 2021]. Unlike linear models, which may struggle to capture complex patterns, 1D-CNNs excel in detecting subtle temporal dependencies and intricate relationships across features over time [Kiranyaz et al., 2019]. This is especially useful for analyzing sequential data, where the ability to recognize patterns evolving over time is essential. Furthermore, 1D-CNNs' capacity to learn localized patterns within sequences enables them to generalize effectively across

varying contexts [Wang et al., 2017], making them suitable for real-world applications where data variability and complexity are common.

Data Pre-processing. Since feature selection is not required in deep learning models like 1D-CNN, we employed different methods for missing value handling, class imbalance handling, and data standardization and transformation. To handle missing values in our time series data, we used forward filling followed by backward filling within each participant's data, a common imputation technique for time series data [Che et al., 2018]. For class imbalance, we balanced the training set by duplicating instances from the minority class (low performers), ensuring an equal 1:1 ratio between classes. Additionally, we standardized all features and transformed the data into a three-dimensional structure, with dimensions representing participants, days, and features. For further details on these processes, refer to Appendix C.1.3.

Modeling Setup. To reduce computational costs, we split the data into an 80% training set and a 20% testing set. For the 1D-CNN approach pipeline, we used 5-fold Leave-One-Participants-Out Cross-Validation (LOPO-CV) on the training set, replacing the LOSO-CV used in the LR approach pipeline due to the computational cost. The entire modeling process was repeated five times, and the average model performance was reported to mitigate stochastic influences. The best model was selected using only the training data, ensuring no information leakage to the test data set. In contrast to the LR pipeline, where data was aggregated at a weekly level, the 1D-CNN approach maintained the input data as daily time series, allowing the model to capture finer temporal patterns. More details about the 1D-CNN model architecture can be found in Appendix C.1.3.

4.4.3 The MTL-1D-CNN Approach

Design Rationale. Multi-Task Learning (MTL) [Ruder, 2017] enhances models by enabling them to perform multiple related tasks simultaneously. This approach allows a model to learn shared representations across tasks, improving its ability to generalize to new data [Caruana, 1997, 1993]. Additionally, when designed appropriately, MTL can support real-world early prediction settings by leveraging knowledge from the secondary task to refine predictions in the

primary task, thereby enhancing practical implementation. In line with this, we introduce a third modeling approach that extends the 1D-CNN approach with a secondary task—predicting prior term GPA—resulting in the MTL-1D-CNN approach. The pre-processing steps for this approach remain the same as those used for 1D-CNN. Below, we detail the modeling setup for this extension.

Modeling Setup. Our MTL approach extends the 1D-CNN network by adding a secondary task: predicting prior term GPA, which is known to correlate with current academic success [Paschall and Freisthler, 2003]. The primary task, predicting end-of-term GPA, is trained on data from the first week of 2018 and tested on 2019 data, while the secondary task is trained on the combined data from the first week of both years. For the secondary task, we only used prior term GPA labels, which is available before the new term begins, ensuring no data leakage and maintaining the ability to make predictions at the end of week one without needing end-of-term GPAs from the 2019 data. Figure 4.2b visualizes the training and testing process for this approach.

This approach is an extension of hard parameter sharing, where both tasks typically use the same dataset and input format. However, since our tasks use different datasets—training data for the primary task and a combination of training and test data for the secondary task—this creates a challenge. Soft parameter sharing is often used in such cases, where each task has its own model and data, but in our approach, both tasks share the same model while using different datasets. To address the imbalance in sample size between the primary and secondary tasks, we resampled the smaller dataset (from the primary task) to match the size of the larger dataset and continued using hard parameter sharing.

4.5 Model Evaluations

In this part, we first evaluate the three approaches' effectiveness in making early predictions about academic performance. We then assess their explainability, focusing on how easily we can extract interpretable insights into the factors influencing students' academic outcomes. Afterward, we examine the fairness of these approachs, evaluating whether particular student groups

are disproportionately impacted by the predictions. Lastly, we explore these approaches' generalizability, assessing how well they maintain performance when applied to new, unseen data.

4.5.1 Effectiveness in Early Predicting Academic Performance.

Since data and code from the reviewed prior research were inaccessible, we defined and compared our approaches against three baselines. The first baseline, 0R (Zero Rule), naively predicted that all students were high performers, reflecting the majority class. The second baseline, 1R-SVM (One Rule), was trained on a single feature—prior term GPA—using a Support Vector Machine (SVM), which was selected as the best-performing model from a comparison of eight classical machine learning models[2]. Lastly, we re-implemented the Long Short-Term Memory (LSTM) model from prior work [Chen and Cui, 2020], which is the only previous research that evaluated the generalizability of their academic performance prediction model. All approaches were evaluated using seven performance metrics: accuracy, precision, recall, F1 score, and AUC, alongside two metrics tailored for imbalanced data: kappa [McHugh, 2012] and balanced accuracy [Brodersen et al., 2010].

To recap, both our LR and 1D-CNN approaches used the 2018 dataset as an "experimental" set to fine-tune their modeling pipelines, addressing factors such as missing value imputation and class imbalance handling. Once the pipelines were finalized, they were applied to the 2019 dataset without modification, ensuring consistent testing across both years. The MTL-1D-CNN approach, however, differs in its setup, as it utilized both the 2018 and 2019 datasets to train and test two tasks, focusing primarily on improving generalizability. Given this distinction, we focus our performance evaluation on the LR and 1D-CNN approaches.

Evaluation Results. As shown in Table 4.2, both the LR and 1D-CNN approaches demonstrated high performance in predicting academic outcomes, with similar results observed across both 2018 and 2019 data, based on information collected no later than the first week of the Spring term. Our results from the 1D-CNN approach are comparable to the previous earliest

[2]Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor [Aha et al., 1991], AdaBoost [Freund and Schapire, 1997], Random Forest [Breiman, 2001], XGBoost [Chen and Guestrin, 2016], Gradient Boosting [Friedman, 2001], and Decision Tree.

Table 4.2: Performance of the LR and 1D-CNN approaches on two years of data, with 0R (Zero Rule), 1R-SVM (One Rule), and a re-implemented LSTM model as baselines. Results are sorted by **Balanced accuracy**. The high performance of the two approaches on both the 2018 and 2019 datasets demonstrates that accurate early prediction is *possible*. The highest-performing approach, based on Balanced Accuracy, is highlighted in **bold**.

Year	Earliest predictable time	Approach	Accuracy	Precision	Recall	F1	AUC	Kappa	Balanced accuracy
2018	any time in Spring term	0R (Zero Rule)	0.771	0.771	1.000	0.871	0.500	0.000	0.500
2018	wk1 in Spring term	LSTM	0.737	0.833	0.769	0.800	0.417	0.417	0.718
2018	before Spring term	1R-SVM (One Rule)	0.766	0.955	0.731	0.828	0.834	0.481	0.807
2018	wk1 in Spring term	LR (Our Approach)	0.915	0.901	**1.000**	0.948	0.962	0.722	0.814
2018	wk1 in Spring term	**1D-CNN (Our Approach)**	**0.948**	**0.958**	0.975	**0.966**	**0.987**	**0.852**	**0.918**
2019	wk1 in Spring term	LSTM	0.611	0.769	0.714	0.741	0.482	-0.033	0.482
2019	any time in Spring term	0R (Zero Rule)	0.679	0.679	1.000	0.809	0.500	0.000	0.500
2019	before Spring term	1R-SVM (One Rule)	0.668	0.815	0.662	0.730	0.682	0.312	0.672
2019	wk1 in Spring term	LR (Our Approach)	0.893	0.894	**0.955**	0.924	0.796	0.745	0.858
2019	wk1 in Spring term	**1D-CNN (Our Approach)**	**0.898**	**0.901**	0.955	**0.927**	**0.866**	**0.758**	**0.866**

prediction study [Sano et al., 2015] that used data collected over the first four weeks of the term, achieving a similar average accuracy of 92% but with predictions made *three weeks earlier*. Both approaches outperformed the 0R, 1R-SVM, and LSTM baselines.

The robustness of both the LR and 1D-CNN approaches across the 2018 and 2019 datasets provides strong evidence that early predictions of student performance, using data available no later than the first week of the term, are *possible*. However, as emphasized earlier, predictive models designed for real-world applications in academic performance must also account for their social and ethical implications. In the following sections, we evaluate the explainability, fairness, and generalizability of all three approaches.

4.5.2 Explainability Evaluation.

We assess the explainability of the three approaches by examining how effectively they provide information that is interpretable and useful in understanding academic performance, especially their potential to offer actionable insights.

Evaluation Results. By its nature, the LR approach offers high interpretability, as it directly maps features to outcomes through easily interpretable coefficients. To aid in understanding

which features most influenced academic performance predictions, we applied feature importance ranking [Bhatt et al., 2020]. Given that feature selection in the LR pipeline occurred iteratively during LOSO-CV, each selected feature was assigned multiple importance scores across iterations. We summed these importance scores to generate final rankings for each feature. In total, the model selected 49 features for the 2018 dataset and 35 features for the 2019 dataset, associated with end-of-term GPA in the first week of the Spring term. We report the top 30 features based on their final importance scores for each year, as shown in Tables 4.3, 4.4.

An example to illustrate reading these tables: the top-ranked feature in Table 4.3 (first row) represents the change in *number of unlock screen events per minute* (in the **Feature** column) at night (in the **Epoch** column) during the second half of the week (in the **Behavioral Change Indicator** column), with the strongest association with end-of-term GPA (in the **Rank** column). Specifically, higher variation in screen unlocks during this period positively impacts a student's GPA (in the **Impact on GPA** column).

In contrast to the LR approach, the 1D-CNN and MTL-1D-CNN approaches, which are based on deep learning techniques, capture more intricate and complex patterns in the data. This capability allows them to model the sequential and temporal dependencies that are essential in behavioral data. However, this comes at the cost of reduced transparency, which makes it challenging to interpret the specific contribution of individual features to the predictions. This "black box" nature of deep learning models introduces difficulties in explaining how certain behaviors influence the outcomes, thereby complicating the interpretability and explainability of these approaches.

Behavioral Patterns Associated with Academic Performance. We analyzed key in-class and outside-classroom behavioral patterns and self-reported factors associated with academic performance by grouping the top features in Tables 4.3, 4.4. Interestingly, we observed a greater number of relevant outside-classroom behaviors than in-class behaviors. Additionally, behavioral shifts frequently occurred on Thursdays in both years (see **Bkp.** column in Tables 4.3, 4.4, indicating that students' weekend behaviors may begin on Fridays rather than Saturdays for a considerable portion of the population. For this analysis, we thus distinguish between "weekday"

Table 4.3: Top 30 selected features in the first week of the 2018 Spring term. The **Feature** column lists the top 30 features selected by the LR approach, ranked by their importance in the **Rank** column. The **Impact on GPA** column shows each feature's weight, indicating its influence on GPA prediction: (+) for a positive association and (-) for a negative one. The **Agg.** column specifies the statistical metric (e.g., average, standard deviation) used to compute each feature within an **Epoch**. The **Behavioral Change Indicator** column identifies if the feature represents a behavioral change associated with GPA, detailing when this change occurs (e.g., in the first or second half of the week, across the full week, or around a breakpoint). The **Bkp.** column marks the specific day indicating a change in weekly behavioral trends, with additional slopes noted for changes before and after the breakpoint, where relevant.

	Feature	Agg.	Epoch	Behavioral Change Indicator	Bkp.	Impact on GPA	Rank
Passive behavioral data	num of unlock screen events per minute		night	second-half slope		1.941 (+)	1
	num of steps taken per bout among all active bouts	min	night	slope all		1.414 (-)	2
	duration of phone remaining unlocked	min	evening	breakpoint	Wed	1.217 (-)	5
	shortest duration of staying awake	avg	24hr			1.205 (+)	6
	duration of all phone interactions	sum	night	slope before breakpoint	Th	1.181 (-)	7
	duration of sedentary bouts	std	24hr	breakpoint	Th	1.103 (+)	8
	time spent at living places	avg	24hr	slope after breakpoint	Th	1.028 (-)	9
	duration of awake		24hr	breakpoint	Th	1.014 (-)	10
	time spent at third-ranked location cluster		24hr	slope before breakpoint	Wed	0.992 (-)	11
	percentage of time in class	avg	24hr			0.952 (+)	14
	num of scans of bluetooth devices owned by others		evening	second-half slope		0.940 (+)	15
	duration of restless	max	24hr	slope after breakpoint	Th	0.889 (-)	18
	time spent at local location clusters	std	morning	second-half slope		0.883 (-)	19
	regularity in circadian movement	std	evening	second-half slope		0.878 (+)	20
	time spent at exercise places in minutes		24hr	first-half slope		0.831 (+)	22
	normalized entropy of local location clusters		24hr	slope after breakpoint	Th	0.829 (+)	23
	time spent at first-ranked location cluster		evening	slope after breakpoint	Th	0.819 (-)	24
	time spent at living places	min	24hr	slope after breakpoint	Wed	0.818 (-)	25
	num of active bouts	std	morning			0.807 (-)	26
	time spent at local location clusters	max	morning	breakpoint	Th	0.801 (+)	27
	time spent at second-ranked location cluster		morning	slope after breakpoint	Th	0.784 (-)	28
	duration of time spent at living places in minutes		24hr	second-half slope		0.775 (-)	29
	time spent at second-ranked location cluster		morning	breakpoint	Th	0.750 (-)	30
Self-report	got lower grades than expected					1.319 (-)	3
	type of service provider					1.315 (-)	4
	experienced helplessness in a difficult situation					0.982 (-)	12
	had problems with partners					0.962 (-)	13
	how unhappy if actual GPA is lower than expected					0.913 (-)	16
	understand less than others about their school					0.896 (-)	17
	felt weak all over					0.850 (-)	21

and "weekend" behaviors starting on Fridays. Throughout this paper, features are referenced by year and rank for consistency. For example, 2018-R14 (*percentage of time in class*) refers

Table 4.4: Top 30 selected features in the first week of the 2019 Spring term.

	Feature	Agg.	Epoch	Behavioral Change Indicator	Bkp.	Impact on GPA	Rank
Passive behavioral data	percentage of time spent at outlier locations		night	slope before breakpoint	Fri	1.502 (-)	1
	num of location bouts at greens		evening	slope before breakpoint	Th	1.271 (+)	2
	num of location bouts with duration >= 10 mins at living places		evening	slope before breakpoint	Th	1.203 (+)	3
	num of location bouts with duration >= 30 mins at exercise places		evening	second-half slope		1.138 (-)	4
	num of location bouts at exercise places	avg	evening	second-half slope		1.030 (-)	5
	num of location bouts with duration >= 30 mins at living places		24hr	second-half slope		1.028 (-)	6
	time spent at second-ranked global location cluster		24hr	slope after breakpoint	Fri	0.959 (-)	7
	time spent at second-ranked location cluster		night	breakpoint	Fri	0.959 (-)	8
	party duration in minutes		night	breakpoint	Fri	0.949 (-)	9
	num of scans of all self-owned bluetooth devices	std	afternoon	first-half slope		0.937 (-)	10
	duration of location bouts at living places	min	morning	second-half slope		0.897 (-)	11
	num of location bouts with duration >= 10 mins at food places		evening	second-half slope		0.864 (+)	12
	duration of location bouts at exercise places	std	morning	breakpoint	Th	0.849 (+)	13
	normalized entropy of local location clusters		morning	second-half slope		0.798 (-)	14
	duration of location bouts at Greek houses	max	night	slope after breakpoint	Fri	0.787 (+)	15
	duration of location bouts at living places	max	night	first-half slope		0.765 (+)	16
	percentage of time spent at greens		afternoon	slope before breakpoint	Th	0.764 (+)	17
	duration of awake in minutes	avg	24hr			0.740 (-)	18
	num of scans of all self-owned bluetooth devices	avg	morning	first-half slope		0.734 (-)	19
	duration of phone remaining unlocked	min	afternoon	breakpoint	Th	0.727 (-)	20
	percentage of time spent near home (within 10m)		morning	first-half slope		0.708 (-)	21
	party duration in minutes	avg	night			0.692 (-)	22
	num of scans of least frequently scanned bluetooth device of others		24hr	breakpoint	Fri	0.683 (-)	23
	num of location bouts at food places		evening	slope after breakpoint	Th	0.682 (+)	24
	num of location bouts with duration >= 10 mins outside		evening	slope all		0.636 (+)	26
	percentage of time spent near home (within 100m)		24hr	breakpoint	Th	0.626 (+)	27
	num of location bouts with duration >= 30 mins at exercise places		evening	slope all		0.620 (-)	28
	time spent at second-ranked cluster in minutes		evening	slope before breakpoint	Th	0.605 (-)	30
Self-report	had trouble sleeping because of pain					0.640 (-)	25
	had traumatic experiences					0.619 (-)	29

to a feature ranked 14th in 2018 (Table 4.3). While associations noted here are not causal, we summarize the observed academic-related behavioral patterns below, along with implications for early prediction. Future research should continue to investigate these patterns to clarify their role in early academic intervention.

Among in-class behaviors, class attendance during the first week of the Spring term is associated with end-of-term GPA, showing that higher attendance correlates positively with academic outcomes (2018-R14). This emphasizes the importance of early engagement in academic activities and suggests that supporting students in establishing consistent attendance patterns could be an effective early intervention strategy. Interestingly, although study duration and study focus time were included in the training process, these factors did not appear among the top

predictors. This absence suggests that attendance might capture a broader engagement factor, while study-specific metrics may require more context or extended observation to reveal their impact on academic performance.

Among outside-classroom behaviors, several patterns were significantly associated with end-of-term GPA. For instance, phone usage shows contrasting effects depending on timing: weekday phone use is negatively associated with GPA (2018-R7, 2019-R19), possibly reflecting distractions during school times, while phone use on weekends shows a positive association with GPA (2018-R1, 2018-R15), perhaps serving as a way for students to unwind after the week. Similarly, time spent at exercise locations during weekdays positively correlates with academic performance (2018-R22), aligning with findings that physical activity supports academic performance [Al-Drees et al., 2016], but this association turns negative when exercise occurs in the evenings on weekends (2019-R4, 2019-R5), possibly suggesting that late-weekend exercise may disrupt academic focus for the upcoming week. Sleep patterns are also crucial; poor quality, frequent wakefulness, and all-nighters predict lower GPAs (2018-R18, 2019-R18), consistent with research linking sleep quality to academic success. Notably, short wakefulness periods positively associate with GPA (2018-R6), a finding that warrants further exploration as it contrasts with studies emphasizing sleep consistency and quality [Okano et al., 2019].

In addition to behaviors, several serious self-reported stressors, including relationship issues (2018-R13), health concerns (2018-R21, 2019-R25), and traumatic experiences (2019-R29), were strongly linked to lower GPAs. Academic-related stressors like underperforming in a prior term (2018-R3) or achieving a lower GPA than expected (2018-R16) were also associated with academic decline, aligning with previous studies on the negative impact of stress on academic outcomes [De Luca et al., 2016; Pereira et al., 2018]. These findings suggest that early mental health support and academic counseling at the start of the term could help students better manage stress and improve resilience, benefiting their academic performance. For a deeper exploration of these patterns, refer to Appendix C.1.4.

4.5.3 Fairness Evaluation.

We assessed fairness, specifically algorithmic bias, across four protected traits: race, first-generation status, gender, and sexual orientation. To evaluate whether the three approaches exhibit biases against these protected groups, we employed three widely used fairness metrics: First, *demographic parity*, which requires that the likelihood of a positive prediction is the same across protected and unprotected groups. Second, *equalized odds*, which ensures that both groups have equal true positive and false positive rates. This metric addresses the limitation of demographic parity, where even a fully accurate classifier may be seen as biased if the actual ratio of positive outcomes differs between groups. Third, *equal opportunity*, a less strict measure than equalized odds, which only requires that the true positive rates be equal across groups. The fairness evaluation was carried out using *Fairlearn* [Bird et al., 2020], a Python toolkit designed to assess and mitigate bias in machine learning models.

While much of the existing literature offers theoretical frameworks for fairness, practical guidelines on acceptable bias thresholds are less established. One exception is demographic parity, where a difference between -0.1 and 0.1, and a ratio between 0.8 and 1.2, is considered *reasonable* [Feldman et al., 2015; Kobayashi and Nakao, 2021; Pessach and Shmueli, 2022]. We extended these thresholds to our assessments of equalized odds and equal opportunity. Given that the 2018 dataset served primarily as an "experimental" dataset for refining our modeling pipelines, our fairness assessment focused on all three approaches on the 2019 dataset. Figure 4.3 visualizes the fairness of each of the three approaches for each group, with values within the *reasonable* range defined above highlighted in light yellow.

Evaluation Results Based on Demographic Parity. Both the LR and 1D-CNN approaches demonstrate generally fair performance for race, gender, and sexual orientation, with small differences and ratios close to 1, which are considered within a reasonable range for fairness. However, for first-generation status, these two approaches show larger differences, suggesting potential biases against this group. In contrast, the MTL-1D-CNN approach consistently shows larger differences for first-generation and sexual orientation, indicating that this approach might have more fairness issues for these traits.

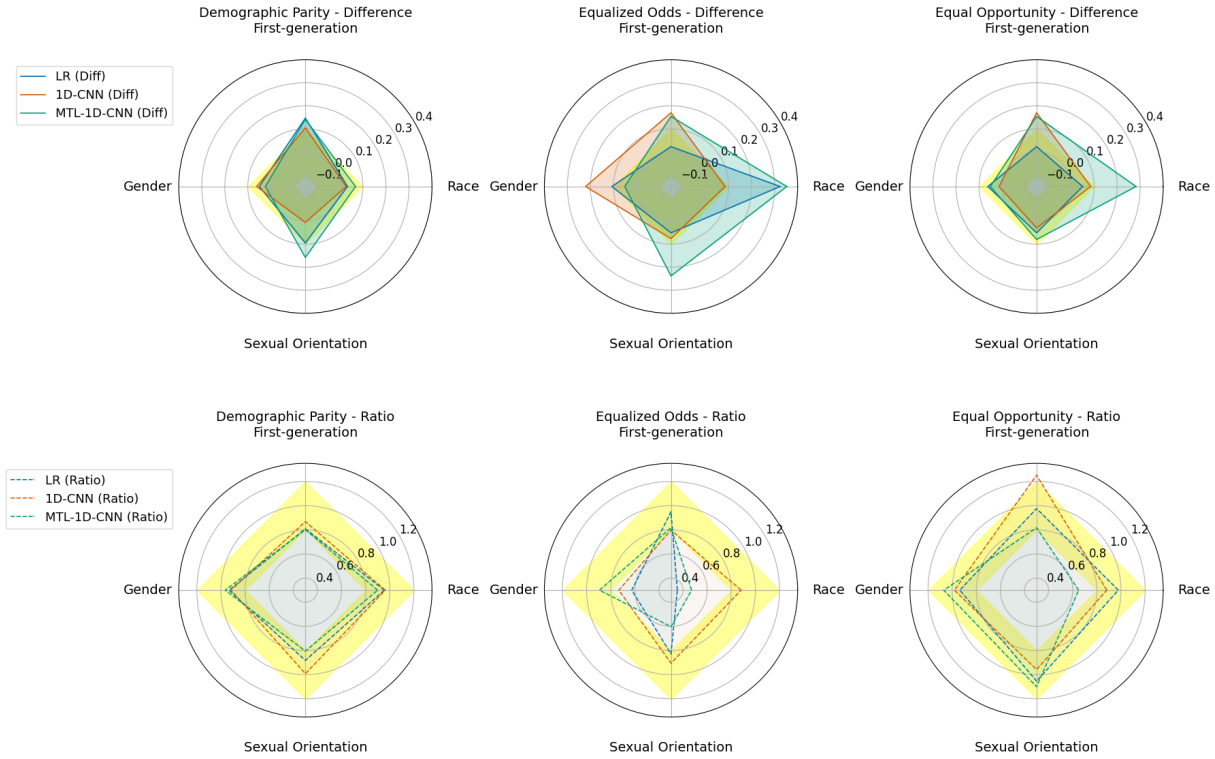


Figure 4.3: Radar charts comparing the fairness performance of three approaches (LR, 1D-CNN, MTL-1D-CNN) across four protected traits (race, first-generation, gender, sexual orientation) using three fairness metrics: demographic parity, equalized odds, and equal opportunity. The first row shows the difference between the protected traits, where the light yellow shaded regions indicate values between -0.1 and 0.1, representing a *reasonable* fair difference. The second row shows the ratio, where the light yellow shaded regions highlight ratio values between 0.8 and 1.2, indicating a *reasonable* fair performance.

Evaluation Results Based on Equalized Odds. The 1D-CNN approach again demonstrates reasonably fair performance for race and sexual orientation, with differences below 0.1 and ratios close to or above 0.9, suggesting that this approach predicts these traits fairly. However, for first-generation status and gender, the differences and ratios are less favorable, indicating potential fairness issues for these groups. The LR approach shows a similar pattern, with relatively better fairness for sexual orientation and first-generation status, but lower fairness for race and gender. In contrast, the MTL-1D-CNN approach, while performing better for gen-

der, displays the largest differences and lowest ratios across most other traits, emphasizing more substantial biases in its predictions.

Evaluation Results Based on Equal Opportunity. The LR approach demonstrates relatively fair treatment for all protected traits, with differences below 0.1 and ratios close to 1. The 1D-CNN approach also shows good performance for race, gender, and sexual orientation, with ratios approaching 1. However, MTL-1D-CNN continues to exhibit larger differences and lower ratios for race and first-generation, suggesting fairness issues for these traits.

4.5.4 Generalizability Evaluation.

Previously, we demonstrated that when training and testing an approach, including data preprocessing and modeling, using data from different students within the same year (a concept referred to as *pipeline-level* generalizability [Xu et al., 2019b]), the LR and 1D-CNN approaches showed robust performance. However, for real-world applications, it is essential to assess whether a pre-trained model can generalize across different contexts (a concept reffered to as *model-level* generalizability [Xu et al., 2019b]), such as applying the model to data from a different year or institution in academic performance prediction. In addition, for real-world academic performance prediction, the most useful models are those that can accurately predict not only students who consistently perform well or poorly but also those who may experience significant changes in performance, such as transitioning from high to low performers.

To assess these aspects of generalizability, we tested our three approaches in two ways. First, we evaluated *model-level* generalizability by training the LR and 1D-CNN approaches on 2018 data and testing them on unseen 2019 data. The MTL-1D-CNN approach, by its design, was assessed on both the 2018 and 2019 datasets, leveraging the multi-task learning framework to test its generalization ability across different years. Second, we analyzed the approaches' accuracy in making predictions for students whose performance remained stable from the Winter term to the following Spring term (i.e., consistently high or low performers) and those whose performance shifted (i.e., transitioning from high to low performers, or vice versa). Specifically, we categorized students into four categories: those who remained high performers (111 participants), those

who remained low performers (29 participants), those who improved to high performers (22 participants), and those who declined to low performers (34 participants). For each approach, we then calculated the percentage of accurately predicted outcomes within these four categories.

Table 4.5: Performance of approaches trained on 2018 data and tested on 2019 data compared to three baselines (seperated by the dashline) to predict end-of-term GPA. Results are sorted by **Balanced accuracy**. The results can indicate *model-level* generalizability of each approach. The MTL-1D-CNN approach significantly outperformed the LSTM baseline. The highest-performing approach, based on Balanced Accuracy, is highlighted in **bold**.

Approach	Accuracy	Precision	Recall	F1	AUC	Kappa	Balanced accuracy
0R (Zero Rule)	0.679	0.679	1.000	0.809	0.500	0.000	0.500
LSTM ([Chen and Cui, 2020])	0.633	0.719	0.752	0.735	0.566	0.136	0.566
1R-SVM (One Rule)	0.668	0.815	0.662	0.730	0.677	0.312	0.672
LR (Our Approach)	0.679	0.679	1.000	0.809	0.652	0.000	0.500
1D-CNN (Our Approach)	0.673	0.732	0.820	0.773	0.592	0.592	0.559
MTL-1D-CNN (Our Approach)	**0.745**	**0.817**	**0.805**	**0.811**	**0.712**	**0.420**	**0.712**

4.5.5 Model-Level Generalizability Evaluation Results

Table 4.5 presents the *model-level* generalizability performance of our three approaches across different years, with the MTL-1D-CNN approach achieving the highest scores across all evaluated metrics. However, neither the LR nor 1D-CNN approaches outperformed the three baseline models, highlighting the challenges in achieving strong generalizability across multiple contexts.

4.5.6 Consistency and Transition Evaluation Results

Figure 4.4 presents a comparison of the approaches' performance in predicting both consistent and transitioning student outcomes. For the "Stay as high performers" category, the 0R baseline, which naively predicts all students as high performers, achieved 100% accuracy as expected. Assessing our approaches, all three performed quite well in identifying students who remained high performers, with accuracy rates exceeding 90%. In the "Stay as low performers" category, all three approaches outperformed the LSTM and 0R baselines, with the MTL-1D-CNN ap-

proach achieving an accuracy above 80%. Interestingly, the prior term GPA (1R-SVM) baseline was especially effective for this group, achieving 100% accuracy, indicating that prior academic performance serves as a strong predictor for students who consistently perform poorly.

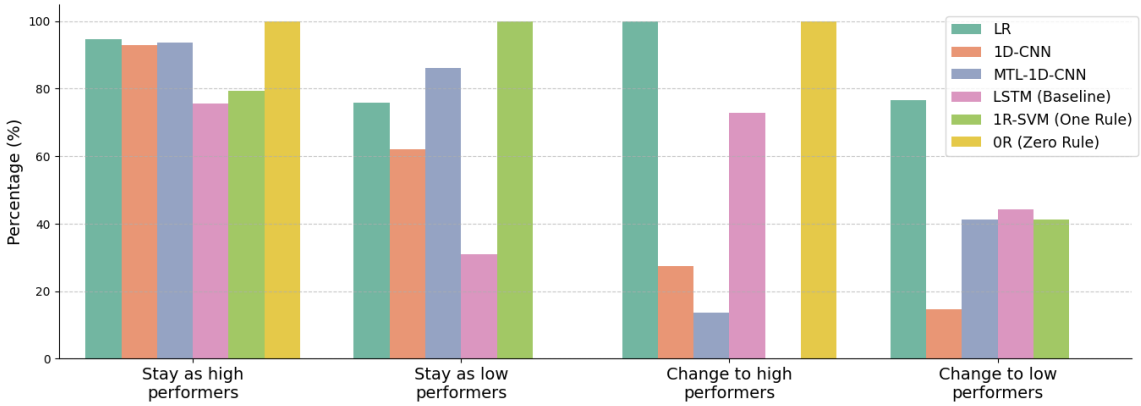


Figure 4.4: Accuracy of three approaches as well as the baselines in predicting academic performance consistency and transitions. It presents the percentage of each approach accurately predicting four categories: remained a high performer, remained a low performer, improved from low to high performer, and declined from high to low performer.

For the "Change to high performers" category, the 0R baseline once again achieved perfect accuracy, by its design. Surprisingly, the LR approach also achieved perfect accuracy in this category, demonstrating its effectiveness in identifying students who improved their performance. However, both deep learning approaches, 1D-CNN and MTL-1D-CNN, showed weaker performance, with accuracy below 35%. When predicting students who transitioned from high to low performance, the LR approach performed the best, with accuracy above 75%, indicating its strength in detecting drops in performance. However, the deep learning approaches and the baselines performed less effectively, with most accuracies around 40%, reinforcing our earlier observations that deep learning models struggle with predicting performance transitions compared to simpler models.

4.6 Discussion, Limitation & Future Work

Our study examines how predictive models can be integrated into collaborative student support systems. Effective academic performance prediction is not just a technical challenge but a socio-technical problem that involves multiple stakeholders. Below, we discuss the insights gained from developing and evaluating early academic performance prediction models, focusing on the importance of human-centered considerations, the trade-offs involved in balancing these aspects, and the scenarios where each approach may be most beneficial for different stakeholders. We also highlight the need for data that supports short-term interventions, summarizing insights learned from analyzing behavioral patterns along with their implications for early interventions. Following this, we discuss the technical insights gained from our study. Finally, we examine the broader implications of our findings and propose future directions for designing predictive models in collaborative student support systems.

Human-Centered Considerations and Trade-offs in Academic Performance Prediction Models. The development of academic performance prediction models requires careful attention to human-centered considerations, particularly their social and ethical implications, to ensure successful implementation in real-world educational settings. Addressing key aspects such as explainability, fairness, and generalizability is essential for fostering trust and usability among key stakeholders, including educators, students, and policymakers [Caruana, 1997; Doshi-Velez and Kim, 2017; Holstein et al., 2019b; Ehsan and Riedl, 2020; Mehrabi et al., 2021; Xu et al., 2023]. While our initial goal was to develop a single model that could balance these three aspects, we found that achieving such a balance is more complex than anticipated. To better understand the inherent trade-offs, we explored three approaches (LR, 1D-CNN, and MTL-1D-CNN) using existing ML/DL techniques.

The LR approach offers the highest level of explainability and demonstrates reasonable fairness, particularly for protected groups such as sexual minorities and first-generation college students. However, its generalizability shows mixed results across different contexts. It performs reliably when predicting students whose academic performance remains stable or changes significantly from one term to the next, especially in identifying students who experience a significant

drop in performance (high performers transitioning to low performers). However, its performance declines when applied to unseen data—a common requirement in real-world applications. This approach may be most beneficial in scenarios where interpretability and fairness are prioritized over generalizability, such as single-institution studies identifying actionable early intervention factors. Its transparency enables students and educators to recognize relevant academic behaviors, supporting proactive interventions, while its fairness helps ensure interventions are designed equitably, benefiting diverse student groups without unintended bias.

The 1D-CNN approach also demonstrates reasonable fairness, but its fairness benefits different protected groups compared to the LR approach. When predicting academic performance within the same dataset, it achieves the highest prediction performance, with an average balanced accuracy of 89.2% across 2018 and 2019. However, this approach struggles with generalizability, showing the worst performance when predicting both consistent and transitioning students and performing comparably worse than the LR approach when applied to new, unseen data. Additionally, its reliance on deep learning techniques reduces explainability, making it more difficult to interpret the influence of individual features on its predictions. Given its strengths and limitations, the 1D-CNN approach may be most suitable in scenarios that prioritize high prediction accuracy within a single, well-defined dataset, particularly when explainability and adaptability are not the primary concern.

The MTL-1D-CNN approach demonstrates the highest generalizability among the three approaches, effectively adapting to new, unseen datasets and consistently predicting students experiencing sustained academic challenges across terms. However, similar to the 1D-CNN approach, its reliance on deep learning reduces explainability, limiting insights into specific behavioral factors influencing predictions. Furthermore, MTL-1D-CNN shows the lowest fairness performance, raising equity concerns for diverse student demographics. Given its strengths and limitations, the MTL-1D-CNN approach may be especially suited to broad, scalable applications where adaptability is crucial, such as multi-institution or multi-term deployments.

Behavioral Data and Early Intervention Implications. As briefly discussed in Section 4.5, while some prior work has provided insights into factors such as personality traits [Wang

et al., 2015a] and students' interactions with OLS that contribute to academic performance [Bravo-Agapito et al., 2021; Waheed et al., 2023], these factors are often difficult to address through short-term interventions. Traits like personality tend to be stable over time, making them less actionable for immediate intervention efforts aimed at improving academic outcomes. Similarly, online behaviors such as visits to forums or number of attempts to access questionnaires, which are collected over extended periods, are challenging to intervene in without a deeper understanding of the motivations driving these behaviors. Therefore, focusing on more dynamic and modifiable factors, such as daily behaviors, could offer greater potential for timely and effective interventions.

The broader HCI community has leveraged passively sensed behavioral data to predict student academic performance and provide insights into daily academic-related behaviors [Wang et al., 2015a]. However, previous research has generally relied on data collected across the entire term, which can delay the timeliness of interventions. In contrast, our work focuses on early-term data to identify key in-class and outside-classroom behavioral patterns, as well as self-reported factors, associated with student academic performance. Our findings reveal that a greater proportion of academic-related factors are linked to outside-classroom behaviors, highlighting the importance of capturing broader aspects of student life. Below, we further discuss the insights gained from these behavioral patterns and propose additional considerations for designing early interventions.

As described in Section 4.5.2, we observed that students' academic-related weekend behaviors tend to begin on Fridays rather than Saturdays, suggesting that routines are generally stable from Monday to Thursday, with shifts starting as early as Friday. Interestingly, this pattern aligns with findings in mood-related studies, where a significant mood shift often occurs on Friday evenings [Stone et al., 2012], signaling a transition into weekend-related behavior and attitudes. This similarity suggests that academic and mood-related routines might both be influenced by the anticipation of the weekend, highlighting Fridays as a potentially critical point for early interventions. Recognizing this shift allows educators and support programs to consider targeted strategies at the end of the week, when students may benefit from reminders to maintain academic habits or engage in well-being activities before the weekend begins.

Additionally, our analysis shows the importance of time granularity in understanding how behaviors impact academic performance. The effects of some behaviors, such as phone usage or time spent exercising, vary depending on the day and time: for instance, phone usage negatively correlates with academic outcomes on weekdays but shows a positive association over the weekend. Similarly, exercise during weekday daytime aligns positively with academic performance, whereas evening exercise on weekends shows a negative relationship. These findings on the timing of behaviors provide critical insights that can enhance short-term interventions by addressing when particular behaviors are more or less beneficial for students. This also highlights the need for collecting continuous data to accurately capture and interpret these time-sensitive patterns.

Technical Insights Learned from Our Experiments. For real-world academic performance prediction applications to be meaningful and ethical, they must address explainability, fairness, and generalizability simultaneously. Our study highlights the complexity of achieving this balance in a single model, underscoring the need for further research in each. Explainability remains a challenge, particularly with DL models that work with high-dimensional behavioral data. Existing interpretability techniques, such as permutation feature importance [Nepal et al., 2022], have shown potential but are computationally costly for datasets with thousands of features. As DL models see broader adoption in educational prediction contexts, advancing methods, such as SHAP [Lundberg, 2017] and LIME [Ribeiro et al., 2016], their interpretability is crucial to enabling actionable insights.

Fairness remains an essential yet under-explored consideration in educational predictive models. Our results show that none of the approaches fully achieves fairness across all protected groups, highlighting the need for effective mitigation strategies. Future work should consider three main types of fairness techniques: pre-processing, which modifies data before model training to reduce bias [Kamiran and Calders, 2012; Feldman et al., 2015]; in-processing, which adjusts the learning algorithm itself to enhance fairness [Zafar et al., 2017; Zhang et al., 2018]; and post-processing, which corrects biases in model predictions after training [Kamiran et al., 2012; Hardt et al., 2016]. Moreover, a critical gap persists in the absence of clear, practical guidelines for validating fairness. Although some efforts have evaluated fairness on real datasets [Awasthi et al.,

2021], concrete standards for what constitutes reasonable fairness are still lacking. Prior work mentions "slack approximations," or ranges of acceptable demographic parity [Delobelle et al., 2021], but more robust, actionable criteria are essential for practical research and deployment.

Generalizability also poses a challenge. Although the ideal goal is to develop a robustly generalizable model, our findings reveal a considerable gap in achieving reliable generalizability. A possible limitation is the exclusion of features closely tied to GPA that were inconsistent across different datasets, either because they were unique to one dataset or completely missing in the other. For instance, *type of service provider* (2018-R4) was excluded from our most generalizable approach (MTL-1D-CNN) as it was not collected in 2019. Lacking a top selected feature like this could significantly affect the model's performance and generalizability, and could be addressed at data collection time without much burden to participants. In addition, introducing other tasks to our MTL approach to either replace the prediction of prior term GPA or to learn more than two tasks in parallel may be of interest for future work, to improve generalizability. Furthermore, we acknowledge the presence of returning students from 2018 to 2019, which may influence result reliability. Future work should consider testing their models/approaches in different contexts to validate trustworthiness and applicability.

Implications for Collaborative Student Support Systems. Predictive models for academic performance are not just technical tools; they function within broader systems where students, educators, advisors, and institutions collaborate to support student success. For these models to be effective in real-world educational settings, they must go beyond accuracy and consider how different stakeholders interpret and act on their outputs. The integration of predictive models into student support systems presents not only technical considerations but also social, material, and theoretical challenges. This study contributes to addressing these challenges by evaluating different modeling approaches, assessing their trade-offs, and exploring how behavioral data can improve early academic interventions.

A central technical challenge in predictive modeling is balancing predictive accuracy with human-centered considerations to ensure that models are both effective and usable. While deep learning models, such as 1D-CNN, achieve higher predictive performance, their complexity re-

duces transparency, making it difficult for stakeholders to interpret and apply the predictions. In contrast, the LR approach offers greater interpretability, allowing students and educators to understand how different factors contribute to predictions, but it struggles with generalizability when applied to new data. This trade-off between explainability and performance highlights a key technical challenge: how can predictive models be designed to optimize both accuracy and usability in multi-stakeholder decision-making? Future work should explore co-design approaches, where educators and institutional decision-makers are actively involved in defining explainability requirements for predictive models.

The integration of predictive models into educational support systems raises social challenges related to fairness and trust. Our results reveal variations in fairness performance across different student demographics, raising concerns about potential bias in AI-driven decision-making. While fairness-aware interventions exist [Selbst et al., 2019; Zhang et al., 2023a; Aird et al., 2024], they often involve trade-offs between model accuracy and equitable outcomes. Additionally, clear guidance on fairness metrics is still missing [Zhang et al., 2024]. From a CSCW perspective, trust in predictive systems cannot be achieved solely through technical bias mitigation—it requires participatory approaches that involve students, educators, and policymakers in defining fairness criteria. Future research should explore human-in-the-loop approaches, where stakeholders collaborate to evaluate, interpret, and refine predictive insights to ensure they align with institutional values and student needs.

Existing academic prediction models rely heavily on sporadic, classroom-derived data, limiting their ability to capture holistic student behaviors [Lu et al., 2018b; Qu et al., 2019; Sukhbaatar et al., 2019]. Our study demonstrates that behavioral data, including sleep patterns, physical activity, and weekend study routines, can provide valuable early signals of academic performance. The findings suggest that integrating behavioral data allows for earlier and more context-aware interventions than traditional models relying solely on classroom engagement. However, the use of behavioral data introduces challenges regarding privacy, data ethics, and responsible data collection. Future research should explore privacy-preserving machine learning techniques and participatory data governance models to ensure responsible data use in educational settings.

A key theoretical challenge in CSCW is how AI-driven insights can be effectively integrated into human-centered decision-making workflows. Additionally, given that different stakeholders often have varying needs, a broader question arises: how can AI systems align with diverse stakeholder goals? Previously, we discuss the potential use cases for different predictive models—for example, educators may prioritize interpretable models, while institutional administrators may favor generalizable solutions. Future research should foster close collaboration between model builders and stakeholders to ensure predictive models are designed with real-world needs in mind. Additionally, future work should explore collaborative AI frameworks that embed predictive models within student support ecosystems, enabling stakeholders to co-define intervention strategies informed by model predictions.

Chapter 5

Conclusion and Future Work

Here, I first summarize the contributions made throughout this dissertation. Then, I discuss a series of opportunities for future work to explore.

5.1 Summary of Contributions

The series of studies presented in this dissertation highlights the need for behavioral sensing systems that not only achieve strong technical performance, but also operate fairly, transparently, and in ways that are sensitive to context and stakeholder needs. Specifically, this dissertation makes the following contributions:

- A longitudinal and situationally anchored understanding of student life. We conducted a six-year, multi-modal longitudinal study at the University of Washington, capturing behavioral and self-report data from over 1,000 undergraduates, along with a focused mixed-methods study during the COVID-19 pandemic. This work surfaced accessibility barriers, mental health challenges, and adaptation strategies during the abrupt shift to remote learning, providing a rich empirical foundation for designing inclusive educational technologies. (Chapters 2)
- A mixed-method audit of fairness in behavioral sensing. We combined in-depth interviews with 14 researchers and practitioners across five countries with a quantitative fairness evaluation of nine existing behavioral sensing models. This audit revealed 16 risks spanning the entire sensing lifecycle, and resulted in a reflexive fairness framework and actionable guidelines to inform more equitable behavioral sensing practices. Our follow-up evaluations on nine behavioral modeling pipelines validated the framework's utility, while providing

empirical evidence on demographic and domain-specific biases. (Chapter 3)

- Human-centered early academic performance prediction. We developed and evaluated three modeling approaches—Logistic Regression, 1D-CNN, and multitask-learning 1D-CNN—that can identify at-risk students as early as week one of an academic term. These models balance predictive accuracy with considerations of generalizability, interpretability, and fairness, illustrating trade-offs that arise when applying human-centered machine learning principles to educational prediction tasks (Chapter 4).

Together, these contributions advance both the technical and socio-technical dimensions of behavioral sensing, providing methods, frameworks, and empirical evidence for building systems that are as responsible as they are effective.

5.2 Opportunities for Future Work

While the behavioral sensing is promising and advancing, many challenges and real-world problems remain unsolved. For instance, how can we develop single behavioral models that balances various HCML principles, leverage behavioral sensing and existing knowledge to create systems that effectively support human well-being, or utilize AI to address the needs of understudied minority groups? Below, I will briefly explore several promising avenues for future research.

Advancing Responsible Behavioral Modeling. A central challenge in behavioral modeling is its limited portability and trustworthiness in real-world contexts. Addressing this requires not only technical robustness but also explicit alignment with societal values such as fairness, generalizability, transparency, and accountability. As my work has shown [Xu et al., 2023; Zhang et al., 2024, 2025], many gaps remain.

First, practical, domain-sensitive criteria for selecting and applying fairness metrics are still lacking. Without such guidance, practitioners risk misinterpreting bias assessments or failing to detect subtle situational disparities. Second, the design of behavioral models that are both fair and generalizable remains an open research problem—particularly when these goals conflict due to data distribution shifts or context-specific behaviors. A long-term research opportunity lies

in developing adaptive modeling frameworks that can dynamically balance fairness, generalizability, and performance as conditions evolve. Third, while my preliminary work [Zhang et al., 2025] demonstrated the potential for interpretable modeling using Logistic Regression for early academic performance prediction, deep learning approaches—often necessary for complex behavioral data—remain opaque. Advancing explainability in this space will require new methods that generate stakeholder-relevant, context-aware explanations while preserving model accuracy. Such innovations could substantially improve the accountability and deployability of behavioral sensing systems.

Developing Behavioral Sensing Systems for Human Well-Being. Improving human well-being—whether in mental health, education, or other life domains—requires more than behavioral prediction. Current systems often lack deep behavioral understanding, which limits their capacity to inform timely, precise, and ethically sound interventions. My research has identified student behavioral patterns linked to academic performance [Zhang et al., 2025], yet correlational evidence alone cannot fully support actionable decision-making. Future work should therefore move toward uncovering causal relationships between behavior and outcomes, using experimental, quasi-experimental, or causal inference methods. Such insights could enable highly targeted interventions that maximize benefit while minimizing harm. Moreover, broadening the scope beyond university student populations will be crucial for ensuring that behavioral sensing systems address the diverse needs of different life stages, cultural contexts, and socio-economic backgrounds. This expansion would allow us to assess the adaptability of both models and interventions, and to test whether strategies effective in one domain (e.g., academic support) can transfer to others (e.g., workplace productivity, clinical mental health care).

Designing AI Systems to Address Diverse Real-World Needs. The principles and frameworks developed in this dissertation—particularly those addressing fairness, explainability, and context-sensitivity—are broadly applicable beyond the specific domains studied here. Many of the challenges identified, such as data representativeness, situational bias, and stakeholder-aligned explanations, also appear in other AI applications aimed at supporting marginalized or

underserved populations.

Building on my findings with first-generation students, individuals from lower socioeconomic backgrounds, and students with disabilities [Nurius et al., 2023; Zhang et al., 2025], future work can extend these human-centered approaches to design AI systems that address diverse real-world needs. This includes developing predictive and assistive tools that adapt to varying cultural, infrastructural, and policy environments; ensuring that such tools incorporate participatory input from affected communities; and evaluating their long-term social impacts alongside their technical performance. By doing so, the field can move toward AI systems that are not only technically robust, but also equitable, contextually grounded, and responsive to the lived experiences of those they serve.

Bibliography

Gregory D Abowd. What next, ubicomp? celebrating an intellectual disappearing act. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012.

Daron Acemoglu. Harms of ai. Technical report, National Bureau of Economic Research, 2021.

Daniel A Adler, Dror Ben-Zeev, Vincent WS Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR mHealth and uHealth*, 2020.

Daniel A Adler, Emily Tseng, Khatiya C Moon, John Q Young, John M Kane, Emanuel Moss, David C Mohr, and Tanzeem Choudhury. Burnout and the quantified workplace: Tensions around personal sensing interventions for stress in resident physicians. *Proceedings of the ACM on Human-computer Interaction*, 2022a.

Daniel A Adler, Fei Wang, David C Mohr, and Tanzeem Choudhury. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *Plos one*, 2022b.

Daniel A Adler, Caitlin A Stamatis, Jonah Meyerhoff, David C Mohr, Fei Wang, Gabriel J Aranovich, Srijan Sen, and Tanzeem Choudhury. Measuring algorithmic bias to analyze the reliability of ai tools that predict depression risk using smartphone sensed-behavioral data. *npj Mental Health Research*, 2024.

N Adler, J Stewart, with the Psychosocial Research Group, et al. The macarthur scale of subjective social status. 2007. *Psychosocial Research Notebook*, 2016.

David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 1991.

Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, 2020.

Tousif Ahmed, Roberto Hoyle, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Understanding physical safety, security, and privacy concerns of people with visual impairments. *IEEE Internet Computing*, 2017.

Amanda Aird, Paresha Farastu, Joshua Sun, Elena Stefancová, Cassidy All, Amy Voida, Nicholas Mattei, and Robin Burke. Dynamic fairness-aware recommendation through multi-agent social choice. *ACM Transactions on Recommender Systems*, 2024.

Abdulmajeed Al-Drees, Hamza Abdulghani, Mohammad Irshad, Abdulsalam Ali Baqays, Abdulaziz Ali Al-Zhrani, Sulaiman Abdullah Alshammari, and Norah Ibrahim Alturki. Physical activity and academic achievement among the medical students: A cross-sectional study. *Medical teacher*, 2016.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019.

Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Enrico Coiera, and Yvonne Rogers. Introduction to the special issue on human-centred ai in healthcare: Challenges appearing in the wild, 2023.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 2020.

Shervin Assari, Ehsan Moazen-Zadeh, Cleopatra Howard Caldwell, and Marc A Zimmerman. Racial discrimination during adolescence predicts mental health deterioration in adulthood: Gender differences among blacks. *Frontiers in public health*, 2017.

Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, et al. Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023.

American College Health Association et al. American college health association-national college health assessment ii: Reference group executive summary spring 2015. *Hanover, MD: American College Health Association*, 2015.

Christoph Augner and Gerhard W Hacker. Associations between problematic mobile phone use and psychological parameters in young adults. *International journal of public health*, 2012.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 2023.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 2016.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020.

Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 1996a.

Aaron T Beck, Robert A Steer, and Gregory K Brown. Beck depression inventory-ii. *San Antonio*, 1996b.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 1995.

Yoav Benjamini and Daniel Yekutieli. Quantitative trait loci analysis using the false discovery rate. *Genetics*, 2005.

Yoav Benjamini, Ruth Heller, and Daniel Yekutieli. Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2009.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020.

Jeffrey P Bigham and Patrick Carrington. Learning from the Front: People with Disabilities as Early Adopters of AI. In *Proceedings of the 2018 HCIC Human-Computer Interaction Consortium*, 2018.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

Carlos Blanco, Mayumi Okuda, Crystal Wright, Deborah S Hasin, Bridget F Grant, Shang-Min Liu, and Mark Olfson. Mental health of college students and their non–college-attending peers: Results from the national epidemiologic study on alcohol and related conditions. *Archives of general psychiatry*, 2008.

Brianna Blaser and Richard E Ladner. Why is data on disability so hard to collect and understand? In *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, 2020.

Brianna Blaser, Cynthia Bennett, Richard E Ladner, Sheryl E Burgstahler, Jennifer Mankoff, C Frieze, and JL Quesenberry. *Perspectives of women with disabilities in computing*. Cambridge, UK: Cambridge Univ. Press, 2019.

Lawrence D Bobo, Melvin L Oliver, Jr James H Johnson, and Valenzuela Abel Jr. *Prismatic metropolis: inequality in Los Angeles*. Russell Sage Foundation, 2000.

Shahab Boumi and Adan Ernesto Vela. Quantifying the impact of students' semester course load on their academic performance. In *2021 ASEE Virtual Annual Conference Content Access*, 2021.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 2006.

Adrian J Bravo, Margo C Villarosa-Hurlocker, and Matthew R Pearson. College student mental health: An evaluation of the dsm–5 self-rated level 1 cross-cutting symptom measure. *Psychological Assessment*, 2018.

Javier Bravo-Agapito, Sonia J Romero, and Sonia Pamplona. Early prediction of undergraduate student's academic performance in completely online learning: A five-year study. *Computers in Human Behavior*, 2021.

Leo Breiman. Random forests. *Machine learning*, 2001.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, 2010.

MA Brodie, EM Pliner, A Ho, Kalina Li, Z Chen, SC Gandevia, and SR Lord. Big data vs accurate data in health research: large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Medical hypotheses*, 2018.

Francois Buet-Golfouse and Islam Utyagulov. Towards fair unsupervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Maarten Buyl and Tijl De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 2024.

Kristen M Byers, Salma Elsayed-Ali, Ebrima Jarjue, Rie Kamikubo, Kyungjun Lee, Rachel Wood, and Hernisa Kacorri. Reflections on remote learning and teaching of inclusive design in HCI. In *3rd Annual Symposium on HCI Education (EduCHI2021)*, 2021.

Abby Cameron-Standerford, Katherine Menard, Christi Edge, Bethney Bergh, Ashley Shayter, Kristen Smith, and Laura VandenAvond. The phenomenon of moving to online/distance delivery as a result of covid-19: Exploring initial perceptions of higher education faculty at a rural midwestern university. In *Frontiers in Education*, 2020.

Vincent A Campbell, Jamylle A Gilyard, Lisa Sinclair, Tom Sternberg, and June I Kailes. Preparing for and responding to pandemic influenza: Implications for people with disabilities. *American journal of public health*, 2009.

Luca Canzian and Mirco Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015a.

Luca Canzian and Mirco Musolesi. Trajectories of depression: Unobtrusive monitoring of de-

pressive states by means of smartphone mobility traces analysis. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015b.

R Caruana. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, 1993.

Rich Caruana. Multitask learning. *Machine learning*, 1997.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 2019.

Charles S Carver. You want to measure coping but your protocol'too long: Consider the brief cope. *International journal of behavioral medicine*, 1997.

Laetitia Cassells. The effectiveness of early identification of 'at risk'students in higher education institutions. *Assessment & Evaluation in Higher Education*, 2018.

Stevie Chancellor. Toward practices for human-centered machine learning. *Communications of the ACM*, 2023.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 2002.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 2018.

Fu Chen and Ying Cui. Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K Villalba, Janine M Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G Creswell, Jennifer Mankoff, J David Creswell, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive

sensing: A machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2021.

Tina Chou, Anu Asnaani, and Stefan G Hofmann. Perception of racial discrimination and psychopathology across three us ethnic minority groups. *Cultural Diversity and Ethnic Minority Psychology*, 2012.

Gabriele Ciciurkaite, Guadalupe Marquez-Velarde, and Robyn Lewis Brown. Stressors associated with the covid-19 pandemic, disability, and mental health: Considerations from the intermountain west. *Stress and Health*, 2021.

Sheldon Cohen and Harry M Hoberman. Positive events and social supports as buffers of life change stress 1. *Journal of applied social psychology*, 1983.

Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. A global measure of perceived stress. *Journal of health and social behavior*, 1983.

Toshka Coleman, Sarina Till, Jaydon Farao, Londiwe Shandu, Nonkululeko Khuzwayo, Livhuwani Muthelo, Masenyani Mbombi, Mamare Bopape, Alastair van Heerden, Tebogo Mothiba, et al. Reconsidering priorities for digital maternal and child health: community-centered perspectives from south africa. *Proceedings of the ACM on Human-Computer Interaction*, 2023.

Patricia Hill Collins. *Intersectionality as critical social theory.* Duke University Press, 2019.

European Commission. Ethics guidelines for trustworthy ai. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf, 2019.

Victor P Cornet and Richard J Holden. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics*, 2018.

Candace Cortiella and Sheldon H Horowitz. The state of learning disabilities: Facts, trends and emerging issues. *New York: National center for learning disabilities*, 2014.

Reagan G Cox, Lei Zhang, William D Johnson, and Daniel R Bender. Academic performance and substance use: findings from a state survey of public high school students. *Journal of school health*, 2007.

Marcus Credé, Sylvia G Roch, and Urszula M Kieszczynka. Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research*, 2010.

Dartmouth. Studentlife study, 2014.

Vedant Das Swain, Koustuv Saha, Manikanta D. Reddy, Hemang Rajvanshy, Gregory D. Abowd, and Munmun De Choudhury. Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. Semantic gap in predicting mental wellbeing through passive sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022.

Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D Abowd, and Munmun De Choudhury. Algorithmic power or punishment: Information worker perspectives on passive sensing enabled ai phenotyping of performance and wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

Vedant Das Swain, Lan Gao, Abhirup Mondal, Gregory D Abowd, and Munmun De Choudhury. Sensible and sensitive ai for worker wellbeing: Factors that inform adoption and resistance for information workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D Lytras, Farhat Abbas, and Jalal S Alowibdi. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion*, 2017.

Susan M De Luca, Cynthia Franklin, Yan Yueqi, Shannon Johnson, and Chris Brownson. The relationship between suicide ideation, behavioral health, and college academic performance. *Community mental health journal*, 2016.

Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter*, 2021.

Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 2001.

Anind K Dey, Gregory D Abowd, and Daniel Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 2001.

Ed Diener, Derrick Wirtz, William Tov, Chu Kim-Prieto, Dong-won Choi, Shigehiro Oishi, and Robert Biswas-Diener. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social indicators research*, 2010.

Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 2013.

Afsaneh Doryab, Jun-Ki Min, Jason Wiese, John Zimmerman, and Jason I. Hong. Detection of behavior change in people with depression. In *AAAI Workshop: Modern Artificial Intelligence for Health Analytics*, 2014.

Afsaneh Doryab, Prerna Chikarsel, Xinwen Liu, and Anind K Dey. Extraction of behavioral features from smartphone and wearable data. *arXiv preprint arXiv:1812.10394*, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 2019.

Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 2009.

Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, 2020.

Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on Human-Computer Interaction*, 2023.

Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. The who in xai: How ai background shapes perceptions of ai explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

Carolyn Ellis, Tony E Adams, and Arthur P Bochner. Autoethnography: an overview. *Historical social research/Historische sozialforschung*, 2011.

Sheena Erete, Yolanda A Rankin, and Jakita O Thomas. I can't breathe: Reflections from black women in cscw and hci. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996.

Nancy J Evans, Ellen M Broido, Kirsten R Brown, and Autumn K Wilke. *Disability in higher education: A social justice approach*. John Wiley & Sons, 2017.

FairLearn Contributors. Fairlearn metrics package, 2022.

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

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

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

Mireia Felez-Nobrega, Charles H Hillman, Kieran P Dowd, Eva Cirera, and Anna Puig-Ribera. Activpal™ determined sedentary behaviour, physical activity and academic achievement in college students. *Journal of sports sciences*, 2018.

Daniel Darghan Felisoni and Alexandra Strommer Godoi. Cell phone usage and academic performance: An experiment. *Computers & Education*, 2018.

Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2015.

Fitbit Team. Fitbit development: Sleep logs, 2023.

Dunigan Folk, Karynna Okabe-Miyamoto, Elizabeth Dunn, and Sonja Lyubomirsky. *Have introverts or extraverts declined in social connection during the first wave of COVID-19?* PsyArXiv, 2020.

Barbara L Fredrickson. Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, 2000.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 1997.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 2001.

David M Frost. Social stigma and its consequences for the socially stigmatized. *Social and Personality Psychology Compass*, 2011.

Jane Cooley Fruehwirth, Siddhartha Biswas, and Krista M Perreira. The covid-19 pandemic and mental health of first-year college students: Examining the effect of covid-19 stressors using longitudinal data. *PloS one*, 2021.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 2016.

Robert P Gallagher. *National survey of counseling center directors 2005*. The International Association of Counseling Services (IACS), 2006.

Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020.

Nan Gao, Max Marschall, Jane Burry, Simon Watkins, and Flora D Salim. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data*, 2022a.

Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, Kaixin Ji, and Flora D Salim. Individual and group-wise classroom seating experience: Effects on student engagement in different courses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2022b.

Saul Geiser and Maria Veronica Santelices. Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes, 2007.

Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.", 2022.

Fausto Giunchiglia, Mattia Zeni, Elisa Gobbi, Enrico Bignotti, and Ivano Bison. Mobile social media usage and academic performance. *Computers in Human Behavior*, 2018.

Jane L Givens, Thomas K Houston, Benjamin W Van Voorhees, Daniel E Ford, and Lisa A Cooper. Ethnicity and preferences for depression treatment. *General hospital psychiatry*, 2007.

Mark E Glickman, Sowmya R Rao, and Mark R Schultz. False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *Journal of clinical epidemiology*, 2014.

Ana Allen Gomes, José Tavares, and Maria Helena P de Azevedo. Sleep and academic performance in undergraduates: a multi-measure, multi-predictor approach. *Chronobiology international*, 2011.

James J Gross and Oliver P John. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 2003.

Farley Grubb. Does going greek impair undergraduate academic performance? a case study. *American Journal of Economics and Sociology*, 2006.

Catherine Hale, Stef Benstead, Jenny Lyus, Evan Odell, and Anna Ruddock. Energy impairment and disability inclusion: Towards an advocacy movement for energy limiting chronic illness, 2020.

Mark Andrew Hall. *Correlation-based feature selection for machine learning.* University of Waikato Hamilton, 1999.

Dominik Hangartner, Daniel Kopp, and Michael Siegenthaler. Monitoring hiring discrimination through online recruitment platforms. *Nature*, 2021.

Gabriella M. Harari, Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science*, 2016.

Gabriella M Harari, Sandrine R Müller, Min SH Aung, and Peter J Rentfrow. Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

Laura Hawryluck, Wayne L Gold, Susan Robinson, Stephen Pogorski, Sandro Galea, and Rima Styra. Sars control and psychological effects of quarantine, toronto, canada. *Emerging Infectious Diseases*, 2004.

Martin Hlosta, Zdenek Zdrahal, and Jaroslav Zendulka. Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the seventh international learning analytics & knowledge conference*, 2017.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019a.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019b.

Melissa K Holt, David Finkelhor, and Glenda Kaufman Kantor. Multiple victimization experiences of urban elementary school students: Associations with psychosocial functioning and academic performance. *Child abuse & neglect*, 2007.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 2020.

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2021.

Jeremy F. Huckins, Alex W. daSilva, Weichen Wang, Elin Hedlund, Courtney Rogers, Subigya K. Nepal, Jialing Wu, Mikio Obuchi, Eilis I. Murphy, Meghan L. Meyer, Dylan D. Wagner, Paul E. Holtzheimer, and Andrew T. Campbell. Mental Health and Behavior of College Students During the Early Phases of the COVID-19 Pandemic: Longitudinal Smartphone and Ecological Momentary Assessment Study. *Journal of Medical Internet Research*, 2020a.

Jeremy F Huckins, Alex W DaSilva, Weichen Wang, Elin Hedlund, Courtney Rogers, Subigya K Nepal, Jialing Wu, Mikio Obuchi, Eilis I Murphy, Meghan L Meyer, et al. Mental health and behavior of college students during the early phases of the covid-19 pandemic: Longitudinal smartphone and ecological momentary assessment study. *Journal of medical Internet research*, 2020b.

Justin Hunt and Daniel Eisenberg. Mental health problems and help-seeking behavior among college students. *Journal of adolescent health*, 2010.

Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Virginia W Huynh, Que-Lam Huynh, and Mary-Patricia Stein. Not just sticks and stones: Indirect ethnic discrimination leads to greater physiological reactivity. *Cultural Diversity and Ethnic Minority Psychology*, 2017.

Maral Jamalova and M Constantinovits. The comparative study of the relationship between smartphone choice and socio-economic indicators. *Int. J. Mark. Stud*, 2019.

Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 2014.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 2015.

Robert I Kabacoff, Daniel L Segal, Michel Hersen, and Vincent B Van Hasselt. Psychometric properties and diagnostic utility of the beck anxiety inventory and the state-trait anxiety inventory with older adult psychiatric outpatients. *Journal of anxiety disorders*, 1997.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 2012.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, 2012.

Jonathan Kanter, Adam Kuczynski, Max Halvorson, Lily Slater, , and Mai Nguyen. Uw covid-19 response study, 2020.

Saffron Karlsen and James Y Nazroo. Relation between racial discrimination, social class, and health among ethnic minority groups. *American journal of public health*, 2002.

Jacob Merew Katamei and Gedion A Omwono. Intervention strategies to improve students' academic performance in public secondary schools in arid and semi-arid lands in kenya. *Int'l J. Soc. Sci. Stud.*, 2015.

Anna Kawakami, Shreya Chowdhary, Shamsi T Iqbal, Q Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. Sensing wellbeing in the workplace, why and for whom? envisioning impacts with organizational stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 2023.

Mike Kent. Disability and elearning: Opportunities and barriers. *Disability Studies Quarterly*, 2015.

Sarah Ketchen Lipson, S Michael Gaddis, Justin Heinze, Kathryn Beck, and Daniel Eisenberg. Variations in student mental health and treatment utilization across us colleges and universities. *Journal of American College Health*, 2015.

Anupam Khan and Soumya K Ghosh. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies*, 2021.

Mariam Khan, Misja Ilcisin, and Katherine Saxton. Multifactorial discrimination as a fundamental cause of mental health inequities. *International Journal for Equity in Health*, 2017.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

Serkan Kiranyaz, Turker Ince, Osama Abdeljaber, Onur Avci, and Moncef Gabbouj. 1-d convolutional neural networks for signal processing applications. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 2021.

Kenji Kobayashi and Yuri Nakao. One-vs.-one mitigation of intersectional bias: A general method for extending fairness-aware binary classification. In *International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence*, 2021.

Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. An ultra-brief screening scale for anxiety and depression: the phq–4. *Psychosomatics*, 2009.

George D Kuh, Jillian L Kinzie, Jennifer A Buckley, Brian K Bridges, and John C Hayek. *What matters to student success: A review of the literature*. National Postsecondary Education Cooperative Washington, DC, 2006.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, 1982.

Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 2010.

Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*, 2011.

Juan A Lara, David Lizcano, María A Martínez, Juan Pazos, and Teresa Riera. A system for knowledge discovery in e-learning environments within the european higher education area–application to student data from open university of madrid, udima. *Computers & Education*, 2014.

Emily G Lattie, Elizabeth C Adkins, Nathan Winquist, Colleen Stiles-Shields, Q Eileen Wafford, and Andrea K Graham. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. *Journal of medical Internet research*, 2019.

Antoinette M Lee, Josephine GWS Wong, Grainne M McAlonan, Vinci Cheung, Charlton Cheung, Pak C Sham, Chung-Ming Chu, Poon-Chuen Wong, Kenneth WT Tsang, and Siew E Chua. Stress and psychological distress among sars survivors 1 year after the outbreak. *The Canadian Journal of Psychiatry*, 2007.

Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. Human-centered approaches to fair and responsible ai.

In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

Matilde Leonardi, Jerome Bickenbach, Tevfik Bedirhan Ustun, Nenad Kostanjsek, and Somnath Chatterji. The definition of disability: what is in a name? *The Lancet*, 2006.

Andrew Lepp, Jacob E Barkley, and Aryn C Karpinski. The relationship between cell phone use and academic performance in a sample of us college students. *Sage Open*, 2015.

Orly Lipka, Marlyn Khouri, and Michal Shecter-Lerner. University faculty attitudes and knowledge about learning disabilities. *Higher Education Research & Development*, 2020.

Lydia T Liu, Serena Wang, Tolani Britton, and Rediet Abebe. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, 2023.

Javier López Zambrano, Juan Alfonso Lara Torralbo, Cristóbal Romero Morales, et al. Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 2021.

Hong Lu, Jun Yang, Zhigang Liu, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM conference on embedded networked sensor systems*, 2010.

Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018a.

Owen HT Lu, Anna YQ Huang, Jeff CH Huang, Albert JQ Lin, Hiroaki Ogata, and Stephen JH Yang. Applying learning analytics for the early prediction of students' academic performance in blended learning. *Journal of Educational Technology & Society*, 2018b.

Nissa B Lucero, Renea L Beckstrand, Lynn Clark Callister, and Ana C Sanchez Birkhead. Prevalence of postpartum depression among hispanic immigrant women. *Journal of the American Academy of Nurse Practitioners*, 2012.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Dante L Mack, Alex W DaSilva, Courtney Rogers, Elin Hedlund, Eilis I Murphy, Vlado Vojdanovski, Jane Plomp, Weichen Wang, Subigya K Nepal, Paul E Holtzheimer, et al. Mental health and behavior of college students during the covid-19 pandemic: Longitudinal mobile smartphone and ecological momentary assessment study, part ii. *Journal of Medical Internet Research*, 2021.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 1947.

Cameron McCarthy, Nikhilesh Pradhan, Calum Redpath, and Andy Adler. Validation of the empatica e4 wristband. In *2016 IEEE EMBS international student conference (ISC)*, 2016.

Michael E McCullough, Robert A Emmons, and Jo-Ann Tsang. The grateful disposition: a conceptual and empirical topography. *Journal of personality and social psychology*, 2002.

Bruce S McEwen. Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York academy of sciences*, 1998.

Deanna LH McFadden. Health and academic success: A look at the challenges of first-generation community college students. *Journal of the American Association of Nurse Practitioners*, 2016.

Ebony O McGee and Danny B Martin. 'you would not believe what i have to go through to prove my intellectual value!' stereotype management among academically successful black mathematics and engineering students. *American Educational Research Journal*, 2011.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 2012.

Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. Generalization and personalization of mobile sensing-based mood inference models: An analysis of college students in eight countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2023.

Lakmal Meegahapola, Dimitris Spathis, Marios Constantinides, Han Zhang, Sofia Yfantidou, Niels van Berkel, and Anind K Dey. Faircomp: 2nd international workshop on fairness and robustness in machine learning for ubiquitous computing. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2024.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 2021.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 2019.

David C Mohr, Ken R Weingardt, Madhu Reddy, and Stephen M Schueller. Three problems with current digital mental health research... and three things we can do about them. *Psychiatric services*, 2017a.

David C. Mohr, Mi Zhang, and Stephen M. Schueller. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual review of clinical psychology*, 2017b.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Margaret E Morris, Kevin S Kuehn, Jennifer Brown, Paula S Nurius, Han Zhang, Yasaman S Sefidgar, Xuhai Xu, Eve A Riskin, Anind K Dey, Sunny Consolvo, et al. College from home during covid-19: A mixed-methods study of heterogeneous experiences. *PloS one*, 2021.

Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D'Mello, Munmun De Choudhury,

Gregory D Abowd, and Thomas Plötz. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019.

Laura Mullins and Michèle Preyde. The lived experience of students with an invisible disability at a canadian university. *Disability & Society*, 2013.

Ann Murphy, Derek Malenczak, and Mina Ghajar. Identifying challenges and benefits of online education for students with a psychiatric disability. *Journal of Postsecondary Education and Disability*, 2019.

Imani Mwalumbwe and Joel S Mtebe. Using learning analytics to predict students' performance in moodle learning management system: A case of mbeya university of science and technology. *The Electronic Journal of Information Systems in Developing Countries*, 2017.

Kevin L Nadal, Katie E Griffin, Yinglee Wong, Kristin C Davidoff, and Lindsey S Davis. The injurious relationship between racial microaggressions and physical health: Implications for social work. In *Microaggressions and Social Work Research, Practice and Education*. Routledge, 2020.

Abdallah Namoun and Abdullah Alshanqiti. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 2020.

Martin Nemeth, Dmitrii Borkin, and German Michalconok. The comparison of machine-learning methods xgboost and lightgbm to predict energy development. In *Computational Statistics and Mathematical Modeling Methods in Intelligent Systems: Proceedings of 3rd Computational Methods in Systems and Software 2019, Vol. 2 3*, 2019.

Subigya Nepal, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Alex Dasilva, Meghan Meyer, and Andrew Campbell. Covid student study: A year in the life of college students during the covid-19 pandemic through the lens of mobile phone sensing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022.

Subigya K Nepal. *Toward the Integration of Behavioral Sensing and Artificial Intelligence*. PhD thesis, Dartmouth College, 2024.

Christina Neudecker, Nadine Mewes, Anne K Reimers, and Alexander Woll. Exercise interventions in children and adolescents with ADHD: A systematic review. *Journal of attention disorders*, 2019.

Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. A comparative analysis of techniques for predicting academic performance. In *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, 2007.

Paula S Nurius, Dana M Prince, and Anita Rocha. Cumulative disadvantage and youth well-being: A multi-domain examination with life course implications. *Child and Adolescent Social Work Journal*, 2015.

Paula S Nurius, Yasaman S Sefidgar, Kevin S Kuehn, Jake Jung, Han Zhang, Olivia Figueira, Eve A Riskin, Anind K Dey, and Jennifer C Mankoff. Distress among undergraduates: Marginality, stressors and resilience resources. *Journal of American college health*, 2021.

Paula S. Nurius, Yasaman S. Sefidgar, Kevin S. Kuehn, Jake Jung, Han Zhang, Olivia Figueira, Eve A Riskin, Anind K Dey, and Jennifer C Mankoff. Distress among undergraduates: Marginality, stressors and resilience resources. *Journal of American College Health*, 2023.

Opeyemi Ojajuni, Foluso Ayeni, Olagunju Akodu, Femi Ekanoye, Samson Adewole, Timothy Ayo, Sanjay Misra, and Victor Mbarika. Predicting student academic performance using machine learning. In *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part IX 21*, 2021.

Kana Okano, Jakub R Kaczmarzyk, Neha Dave, John DE Gabrieli, and Jeffrey C Grossman. Sleep quality, duration, and consistency are associated with better academic performance in college students. *NPJ science of learning*, 2019.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2019.

Mallie J Paschall and Bridget Freisthler. Does heavy drinking affect academic performance in college? findings from a prospective study of high achievers. *Journal of Studies on Alcohol*, 2003.

Leisi Pei and Hongbin Wu. Does online learning work better than offline learning in undergraduate medical education? a systematic review and meta-analysis. *Medical education online*, 2019.

Juliana L Pereira, Gisela Maria Guedes-Carneiro, Liana R Netto, Patrícia Cavalcanti-Ribeiro, Sidnei Lira, José F Nogueira, Carlos A Teles, Karestan C Koenen, Aline S Sampaio, Lucas C Quarantini, et al. Types of trauma, posttraumatic stress disorder, and academic performance in a population of university students. *The Journal of Nervous and Mental Disease*, 2018.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 2022.

Betty Pfefferbaum and Carol S North. Mental health and the covid-19 pandemic. *New England Journal of Medicine*, 2020.

Amy Phillips, Katherine Terras, Lori Swinney, and Carol Schneweis. Online disability accommodations: Faculty experiences at one public university. *Journal of Postsecondary Education and Disability*, 2012.

Margaret Price. *Mad at school: Rhetorics of mental disability and academic life*. University of Michigan Press, 2011.

Cassidy Pyle, Nicole B Ellison, and Nazanin Andalibi. Social media and college-related social support exchange for first-generation, low-income students: The role of identity disclosures. *Proceedings of the ACM on Human-Computer Interaction*, 2023.

Shaojie Qu, Kan Li, Bo Wu, Xuri Zhang, and Kaihao Zhu. Predicting student performance and deficiency in mastering knowledge points in moocs using multi-task learning. *Entropy*, 2019.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020.

Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 2007.

Rapids. Rapids documentation, V.1.6. URL https://www.rapids.science/1.6/.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

Kathryn E Ringland, Jennifer Nicholas, Rachel Kornfield, Emily G Lattie, David C Mohr, and Madhu Reddy. Understanding mental ill-health as psychosocial disability: Implications for assistive technology. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019.

Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous agents and multi-agent systems*, 2019.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Shanna Russ and Foad Hamidi. Online learning accessibility during the covid-19 pandemic. In *Proceedings of the 18th International Web for All Conference*, 2021.

Daniel W Russell. Ucla loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of personality assessment*, 1996.

Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, 2015a.

Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 2015b.

Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D. Abowd, and Munmun De Choudhury. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2017.

Koustuv Saha, Ted Grover, Stephen M Mattingly, Vedant Das Swain, Pranshu Gupta, Gonzalo J Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. Person-centered predictions of psychological constructs with social media contextualized by multimodal sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021.

Kristy Sanderson and Gavin Andrews. Prevalence and severity of mental health-related disability and relationship to diagnosis. *Psychiatric services*, 2002.

Akane Sano, Andrew J Phillips, Z Yu Amy, Andrew W McHill, Sara Taylor, Natasha Jaques, Charles A Czeisler, Elizabeth B Klerman, and Rosalind W Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2015.

Michael T Schmitt and Nyla R Branscombe. The meaning and consequences of perceived discrimination in disadvantaged and privileged social groups. *European review of social psychology*, 2002.

Yasaman S Sefidgar, Woosuk Seo, Kevin S Kuehn, Tim Althoff, Anne Browning, Eve Riskin, Paula S Nurius, Anind K Dey, and Jennifer Mankoff. Passively-sensed behavioral correlates

of discrimination events in college students. *Proceedings of the ACM on Human-computer Interaction*, 2019a.

Yasaman S. Sefidgar, Woosuk Seo, Kevin S. Kuehn, Tim Althoff, Anne Browning, Eve Riskin, Paula S. Nurius, Anind K. Dey, and Jennifer Mankoff. Passively-sensed behavioral correlates of discrimination events in college students. *Proc. ACM Hum.-Comput. Interact.*, 2019b.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 2019.

Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, 2008.

James M. Shultz. Mental Health Consequences of Infectious Disease Outbreaks, 2020.

Delar K Singh. Educational rights of college students with disabilities. *College Student Journal*, 2019.

Julie Smart. *Disability across the developmental life span: For the rehabilitation counselor.* Springer Publishing Company, 2011.

Bruce W Smith, Jeanne Dalen, Kathryn Wiggins, Erin Tooley, Paulette Christopher, and Jennifer Bernard. The brief resilience scale: assessing the ability to bounce back. *International journal of behavioral medicine*, 2008.

Changwon Son, Sudeep Hegde, Alec Smith, Xiaomei Wang, and Farzan Sasangohar. Effects of covid-19 on college students' mental health in the united states: Interview survey study. *Journal of medical internet research*, 2020.

C. D. Spielberger, R. L. Gorsuch, R. Lushene, P. R. Vagg, and G. A. Jacobs. *Manual for the State-Trait Anxiety Inventory.* Consulting Psychologists Press, 1983.

Michelle J Sternthal, Natalie Slopen, and David R Williams. Racial disparities in health: how much does stress really matter? 1. *Du Bois review: social science research on race*, 2011.

Arthur A Stone, Stefan Schneider, and James K Harter. Day-of-week mood patterns in the united states: On the existence of 'blue monday','thank god it's friday'and weekend effects. *The Journal of Positive Psychology*, 2012.

Kim Storrie, Kathy Ahern, and Anthony Tuckett. A systematic review: students with mental health problems—a growing problem. *International journal of nursing practice*, 2010.

Susan Stuntzner and Michael Hartley. Resilience, coping, & disability: The development of a resilience intervention. *Vistas Online*, 2014.

Otgontsetseg Sukhbaatar, Tsuyoshi Usagawa, and Lodoiravsal Choimaa. An artificial neural network based early prediction of failure-prone students in blended learning course. *International Journal of Emerging Technologies in Learning (iJET)*, 2019.

Evren Sumuer. The effect of mobile phone usage policy on college students' learning. *Journal of Computing in Higher Education*, 2021.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, et al. Human-centered responsible artificial intelligence: Current & future trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

Murtaza Tamjeed, Vinita Tibdewal, Madison Russell, Michael McQuaid, Tae (Tom) Oh, and Kristen Shinohara. Understanding disability services toward improving graduate student support. In *ASSETS '21: The 23rd International ACM SIGACCESS Conference on Computers and Accessibility, Virtual Event, USA, October 18-22, 2021*, 2021.

Andrew J Thayer, Clayton R Cook, Aria E Fiat, Meghanne N Bartlett-Chase, and Jessie M Kember. Wise feedback as a timely intervention for at-risk students transitioning into high school. *School Psychology Review*, 2018.

Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 1975.

Paul D Trapnell and Jennifer D Campbell. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. *Journal of personality and social psychology*, 1999.

Mickey T Trockel, Michael D Barnes, and Dennis L Egget. Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors. *Journal of American college health*, 2000.

Catrine Tudor-Locke, Ho Han, Elroy J Aguiar, Tiago V Barreira, John M Schuna Jr, Minsoo Kang, and David A Rowe. How fast is fast enough? walking cadence (steps/min) as a practical estimate of intensity in adults: a narrative review. *British Journal of Sports Medicine*, 2018.

Civil Service Commission Department o f Labor US Equal Employment Opportunity Commission, Department o f Justice, et al. Uniform guidelines on employee selection procedures. *Federal Register*, 1978.

Julio Vega, Meng Li, Kwesi Aguillera, Nikunj Goel, Echhit Joshi, Kirtiraj Khandekar, Krina C Durica, Abhineeth R Kunta, and Carissa A Low. Reproducible analysis pipeline for data streams: Open-source software to process data collected with mobile devices. *Frontiers in Digital Health*, 2021.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, 2018.

Aleksandar Višnjić, Vladica Veličković, Dušan Sokolović, Miodrag Stanković, Kristijan Mijatović, Miodrag Stojanović, Zoran Milošević, and Olivera Radulović. Relationship between the manner

of mobile phone use and depression, anxiety, and stress in university students. *International journal of environmental research and public health*, 2018.

Hajra Waheed, Saeed-Ul Hassan, Raheel Nawaz, Naif R Aljohani, Guanliang Chen, and Dragan Gasevic. Early prediction of learners at risk in self-paced education: A neural network approach. *Expert Systems with Applications*, 2023.

Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, Steffi Weidt, et al. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth*, 2016.

Gregory M Walton and Geoffrey L Cohen. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, 2007.

Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

R. Wang, F. Chenand Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. 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*, 2014a.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 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*, 2014b.

Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. Smartgpa: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015a.

Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015b.

Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018a.

Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018b.

Weichen Wang, Gabriella M. Harari, Rui Wang, Sandrine R. Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T. Campbell. Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018c.

Xiaomei Wang, Sudeep Hegde, Changwon Son, Bruce Keller, Alec Smith, and Farzan Sasangohar. Investigating mental health of us college students during the covid-19 pandemic: cross-sectional survey study. *Journal of medical Internet research*, 2020.

Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, 2017.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 2009.

David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures

of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 1988.

Frank W Weathers, Brett T Litz, Terence M Keane, Patrick A Palmieri, Brian P Marx, and Paula P Schnurr. The ptsd checklist for dsm-5 (pcl-5). *Scale available from the National Center for PTSD*, 2013.

Frank Wilcoxon. *Individual comparisons by ranking methods*. Springer, 1992.

David R Williams, Yan Yu, James S Jackson, and Norman B Anderson. Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of health psychology*, 1997.

Thomas R Wolanin and Patricia E Steele. Higher education opportunities for students with disabilities: A primer for policymakers. *The Institute for Higher Education Policy*, 2004.

Tammy Wyatt and Sara B Oswalt. Comparing mental health issues among undergraduate and graduate students. *American journal of health education*, 2013.

Qualtric XM. Qualtric xm: The leading experience management software, 2002. URL https://www.qualtrics.com.

Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 2022a.

Xing Xu, Jianzhong Wang, Hao Peng, and Ruilin Wu. Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, 2019a.

Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019b.

Xuhai Xu, Prerna Chikersal, Janine M Dutcher, Yasaman S Sefidgar, Woosuk Seo, Michael J Tumminia, Daniella K Villalba, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, et al. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. *arXiv preprint arXiv:2211.02733*, 2022b.

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2023.

Zikang Xu, Jun Li, Qingsong Yao, Han Li, Xin Shi, and S Kevin Zhou. A survey of fairness in medical image analysis: Concepts, algorithms, evaluations, and challenges. *arXiv preprint arXiv:2209.13177*, 2022c.

Mustafa Yağcı. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 2022.

Huaxiu Yao, Defu Lian, Yi Cao, Yifan Wu, and Tao Zhou. Predicting academic performance for college students: a campus behavior perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2019.

Stephen S Yau and Fariaz Karim. An adaptive middleware for context-sensitive communications for real-time applications in ubiquitous computing environments. *Real-Time Systems*, 2004.

Johnson Yeboah and George Dominic Ewur. The impact of whatsapp messenger usage on students performance in tertiary institutions in ghana. *Journal of Education and practice*, 2014.

Sofia Yfantidou, Marios Constantinides, Dimitris Spathis, Athena Vakali, Daniele Quercia, and Fahim Kawsar. Beyond accuracy: A critical review of fairness in machine learning for mobile and wearable computing. *arXiv preprint arXiv:2303.15585*, 2023a.

Sofia Yfantidou, Pavlos Sermpezis, Athena Vakali, and Ricardo Baeza-Yates. Uncovering bias in personal informatics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2023b.

Dong Whi Yoo, Hayoung Woo, Sachin R Pendse, Nathaniel Young Lu, Michael L Birnbaum, Gregory D Abowd, and Munmun De Choudhury. Missed opportunities for human-centered ai research: Understanding stakeholder collaboration in mental health ai research. *Proceedings of the ACM on Human-Computer Interaction*, 2024.

Liang-Chih Yu, Cheng-Wei Lee, HI Pan, Chih-Yueh Chou, Po-Yao Chao, ZH Chen, SF Tseng, CL Chan, and K Robert Lai. Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning*, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Han Zhang, Margaret E Morris, Paula S Nurius, Kelly Mack, Jennifer Brown, Kevin S Kuehn, Yasaman S Sefidgar, Xuhai Xu, Eve A Riskin, Anind K Dey, et al. Impact of online learning in the context of covid-19 on undergraduates with disabilities and mental health concerns. *ACM Transactions on Accessible Computing (TACCESS)*, 2022.

Han Zhang, Leijie Wang, Yilun Sheng, Xuhai Xu, Jennifer Mankoff, and Anind K Dey. A framework for designing fair ubiquitous computing systems. In *Adjunct Proceedings of the*

2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing, 2023a.

Han Zhang, Leijie Wang, Yilun Sheng, Xuhai Xu, Jennifer Mankoff, and Anind K. Dey. A Framework for Designing Fair Ubiquitous Computing Systems. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, 2023b.

Han Zhang, Vedant Das Swain, Leijie Wang, Nan Gao, Yilun Sheng, Xuhai Xu, Flora D Salim, Koustuv Saha, Anind K Dey, and Jennifer Mankoff. Illuminating the unseen: A framework for designing and mitigating context-induced harms in behavioral sensing. *arXiv preprint arXiv:2404.14665*, 2024.

Han Zhang, Yiyi Ren, Paula S. Nurius, Jennifer Mankoff, and Anind K. Dey. Towards Human-Centered Early Prediction Models for Academic Performance in Real-World Contexts, 2025.

Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, 2018.

Appendix A

Appendix: Capturing and Understanding Student Experience and Needs

A.1 Details for Section 2.3

Table A.1: Description of survey scales.

Scale Name Abbreviation	Short Description	Scoring Range	Year	Collection Time
UCLA Short-form UCLA Loneliness Scale	A 10-item scale measuring one's subjective feelings of loneliness as well as social isolation. Items 2, 6, 10, 11, 13, 14, 16, 18, 19, and 20 of the original scale are included in the short form. Higher values indicate more subjective loneliness.	10 - 40	1,2,3,4	pre, post
Social Fit Sense of Social and Academic Fit Scale	A 17-item scale measuring the sense of social and academic fit of students at the institution where this study was conducted. Higher values indicate higher feelings of belongings.	17 - 119		
2-Way SSS 2-Way Social Support Scale	A 21-item scale measuring social supports from four aspects (a) giving emotional support, (b) giving instrumental support, (c) receiving emotional support, and (d) receiving instrumental support. Higher values indicate more social support.	(a) 0 - 25 (b) 0 - 25 (c) 0 - 35 (d) 0 - 20		
PSS Perceived Stress Scale	A 14-item scale used to assess stress levels during the last month. Note that Year 1 used the 10-item version. Higher values indicate more perceived stress.	0 - 56 (Year 2,3,4) 0 - 40 (Year 1)		

ERQ Emotion Regulation Questionnaire	A 10-item scale assessing individual differences in the habitual use of two emotion regulation strategies: (a) cognitive reappraisal and (b) expressive suppression. Higher scores indicate more habitual use of reappraisal/suppression.	(a) 1 - 7 (b) 1 - 7		
BRS Brief Resilience Scale	A 6-item scale assessing the ability to bounce back or recover from stress. Higher scores indicate more resilient from stress.	1 - 5		
CHIPS Cohen-Hoberman Inventory of Physical Symptoms	A 33-item scale measuring the perceived burden from physical symptoms, and resulting psychological effect during the past 2 weeks. Higher values indicate more perceived burden from physical symptoms.	0 - 132		
STAI State-Trait Anxiety Inventory for Adults	A 20-item scale measuring State-Trait anxiety. Year 1 used the State version, while other years used the Trait version. Higher values indicate higher anxiety.	20 - 80		
CES-D Center for Epidemiologic Studies Depression Scale Cole version	A 10-item scale measuring current level of depressive symptomatology, with emphasis on the affective component, depressed mood. Year 2 used the 9-item version. Higher scores indicate more depressive symptoms.	0 - 30 (Year 1,3,4) 0 - 27 (Year 2)		
BDI2 Beck Depression Inventory-II	A 21-item detect depressive symptoms. Higher values indicate more depressive symptoms. 0-13: minimal to none, 14-19: mild, 20-28: moderate and 26-63: severe.	0 - 63		
MAAS Mindful Attention Awareness Scale	A 15-item scale assessing a core characteristic of mindfulness. Year 1 used a 7-item version, while other years used the full version. Higher values indicate higher mindfulness.	1 - 6		
BFI10 The Big-Five Inventory-10	A 10-item scale measuring the Big Five personality traits Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness. The higher the score, the greater the tendency of the corresponding personality.	1 - 5	1,2,3,4	pre
Brief-COPE Brief Coping Orientation to Problems Experienced	A 28-item scale measuring (a) adaptive and (b) maladaptive ways to cope with a stressful life event. Higher values indicate more effective/ineffective ways to cope with a stressful life event.	(a): 0 - 3 (b): 0 - 3		

2,3,4 pre, post

GQ Gratitude Questionnaire	A 6-item scale assessing individual differences in the proneness to experience gratitude in daily life. Higher scores indicate a greater tendency to experience gratitude.	6 - 42		
FSPWB Flourishing Scale & Psychological Well-Being Scale	An 8-item scale measuring the psychological well-being. Higher scores indicate a person with "more psychological resources and mental strengths".	8 - 56		
EDS Everyday Discrimination Scale	A 9-item scale assessing everyday discrimination. Higher values indicate more frequent experience of discrimination.	0 - 45		
CEDH Chronic Work Discrimination and Harassment	A 12-item scale assessing experiences of discrimination in educational settings. Higher values indicate more frequent experience of discrimination in the work environment.	0 - 60		
B-YAACQ The Brief Young Adult Alcohol Consequences Questionnaire (optional)	A 24-item scale measuring the alcohol problem severity continuum in college students. Higher values indicates more severe alcohol problems.	0 - 24		
PHQ-4 Patient Health Questionnaire 4	A 4-item scale assessing (a) mental health, (b) anxiety, and (c) depression. Higher values indicate higher risk of mental health, anxiety, and depression.	(a): 0 - 12 (b): 0 - 6 (c): 0 - 6	2,3,4	Weekly EMA
PSS-4 Perceived Stress Scale 4	A 4-item scale assessing stress levels during the last month. Higher values indicates more perceived stress.	0 - 16		
PANAS Positive and Negative Affect Schedule	A 10-item scale measuring the level of (a) positive and (b) negative affects. Higher values indicates larger extent.	(a): 0 - 20 (b): 0 - 20		

A.1.1 Implementation of Low-Level Behavior Features

Physical Activity. For a given day/epoch, we counted the number of times a student's activity type changes (e.g., from "still" to "walking"), the number of unique activity types, and the most frequently logged activity type.

Application Usage. We pre-processed the data to exclude system apps from our feature

Table A.2: Passive-sensing data and extracted low-level behavior features.

Source	Sensor	Sampling frequency	Low-level Behavior features
Smartphone	Physical activity	Every minute	Most common activity, number of activities
	Application usage	Event-based	Number of used apps, most commonly used app, most common app category, apps used per minute
	Battery	Event-based	Number of charging sessions, total charging time
	Bluetooth	Every 3 minutes	Number of scans, number of unique devices, number of scans of least frequent device, number of scans of most frequent device, etc.
	Calls	Event-based	Number of incoming calls, number of outgoing calls, number of missed calls, duration of incoming calls, duration of outgoing calls, etc.
	Locations	Every minute	Total distance traveled, time spent at/near home, average traveling speed, percentage of time moving, time spent at top 3 location clusters, etc.
	Location map	Every minute	Time at exercise-labeled places, time at food-labeled places, time at fraternity-labeled places, time at greens-labeled places, time at living-labeled places, time at study-labeled places, etc.
	Screen	Event-based	Sum duration of phone interactions, average duration of phone interactions, standard deviation of interaction durations, time of first unlock event, time of last unlock event, number of unlocks per minute, etc.
	WiFi	Every 3 minutes	Number of unique access points, most frequent access point
Fitbit	Sleep	Every minute	Time in bed, awake duration, asleep duration, restless duration, sleep efficiency, etc.
	Step count	Every minute	Total step count, number of active bouts, average duration of active bouts, average steps per active bout, start time of longest active bout, etc.

computation to focus mainly on user installed applications (UIA). For a given day/epoch, we calculated the number of unique apps used, and the most commonly used app and app category. We also calculated the average number of apps used per minute by the user.

Battery. We calculated the number of times users charge their phones and the total battery charging time to indicate how often and how long users charge their phones.

Bluetooth. We applied the K-means clustering algorithm to scanned Bluetooth addresses based on their frequency in the data set, and grouped the devices into 2 or 3 clusters depending on which can better separate the data points with more concentrated clusters, to differentiate the person's own devices (labeled as "self") and other people's devices (labeled as "others") [Doryab et al., 2018]. We then calculated statistical features for each group of devices, such as the number of scans of most/least frequent device of self/others, the number of unique devices of self/others, total and average number of scans of all devices of self/others, *etc.*

Calls. The call logs provide session information for incoming, outgoing and missed calls. We computed the total number as well as the total duration of calls belonging to each call type.

Locations. We extracted location variance, radius of gyration, total distance traveled and circadian movement features described in [Doryab et al., 2018]. We used DBSCAN [Ester et al., 1996] to group static location samples into clusters, and calculated the statistical features (e.g., sum, mean, standard deviation, maximum, and minimum) on the duration of stay at each cluster. In addition, we calculated the entropy of the duration of stay at each cluster to evaluate how students distributed their time. We inferred students' home locations by clustering their location data at night (12am to 6am). We considered a potential cluster to be a home location if the student stays there for more than 3 days in a row, and the dwelling time at the cluster is at least 80% of each night. We then calculated the total time spent at home (within 10 meters from home) and near home (within 100 meters from home) accordingly based on their home locations.

Location Map. To better map GPS location data to meaningful places, we hand labeled the boundaries of places of interest (i.e., exercise, food, frat house, greens, dorms/living and study) on campus to create a location map. For each location sample, we assigned a map label to it by comparing it against the location map. We then grouped consecutive samples of the same map label into bouts and calculated statistical features on the durations of the bouts.

Screen. We used screen data to define a phone interaction session as a time series with a screen status of "on" at the beginning and a screen status of "off" or "locked" at the end of the session. Similarly, we defined a screen unlock session to be a time series with a screen status of "unlocked" at the beginning and a screen status of "locked" at the end. We then computed statistical features (e.g., sum, max, min, average, standard deviation) on the duration of interaction and unlock sessions. In addition, we extracted the time information of the first and last occurrence of different types of screen events (i.e., on, off, unlock and lock), and calculated the average number of unlocks per minute to indicate the frequency of a user initiating a phone interaction.

WiFi. We counted the number of unique WiFi access points sensed by the phone and identified the most frequently detected access point.

Sleep. We obtained students' sleep logs through the Fitbit API (v1.0), which contain per-minute data of sleep status (i.e., asleep, restless, and awake) throughout each sleep episode. We grouped consecutive sleep samples of the same status into sleep bouts and calculate statistical features on the asleep, restless and awake bouts such as the total number of awake bouts, the start time, end time, max, min and average duration of asleep bouts, restless bouts, *etc.* We also considered Fitbit summary data [Fitbit Team, 2023] as part of our daily features, including duration and efficiency of sleep, time in bed, *etc.*

Step Count. We computed step features from the minute-by-minute data returned by the Fitbit API. Epidemiological studies report a mean daily cadence of 7.7 steps per minute at the population level [Tudor-Locke et al., 2018]. We used 12 steps per minute as a threshold to determine if a person is active or not in that minute. We grouped consecutive active or inactive samples into active or sedentary bouts, and calculated statistical features on the duration and step count of the bouts. We also extracted the start and end time of the active bout with the longest duration and the bout with the most steps.

A.2 Details for Section 2.4

A.2.1 COVID-Specific Survey Questions

M1. Spring Online Classes Concerns/Stress

- (In pre-term survey) Do you have any of the following concerns about classes going online in Spring Quarter? (In post-term survey) How stressful were any of the following experiences due to the change to remote instruction in Spring Quarter?
  - Moving degree requirements to another quarter
  - Delaying graduation due to degree requirements
  - Impact on your visa
  - Impact on your status (e.g., Dean's list)
  - Impact on your admission to major
  - Impact on your financial aid status
  - Having academic requirements that cannot be accomplished online

– Other (please specify)

M2. COVID-19 Related Adversity

- Did quarantine or other effects of COVID-19 add to conflict/tension with household members?
- Did quarantine or other effects of COVID-19 lead you to feel isolated?
- Was someone in your family positive for COVID-19?
- Did someone in your family develop a serious health problem?
- Did a family member or a close friend die?
- Were you positive for COVID-19?
- Were you quarantined because of COVID-19?
- You experienced discrimination that attribute to COVID-19?

Q1. Do you have a medical condition that puts you at risk for complications associated with COVID-19?

A.2.2 Timeline of Announcements and Events of Relevance

Date	Event
Feb 28	Evidence of community spread discovered locally
Feb 29	First COVID-19 related death discovered locally
Mar 6	Announcement that classes would officially switch to online
Mar 13	Last day of instruction for Winter quarter and announcement that Spring quarter will begin online
Mar 18	Announcement that Spring quarter will be fully online
Mar 18	*Earliest date a student took the pre-term survey*
Mar 20	Last day of final exams for Winter quarter
Mar 23	The stay-at-home order was issued
Mar 30	Instruction for Spring quarter begins
Apr 8	*Latest date a student took the pre-term survey*
Jun 5	Last day of instruction for Spring quarter
Jun 7	*Earliest date a student took the post-term survey*
Jun 12	Last day of final exams for Spring quarter
Jun 22	*Latest date a student took the post-term survey*
Jun 2020; Mar 2021	Interviews conducted

A.2.3 Interview Guide

- **Background Questions**
  - Can you tell me your year in school and a little bit about yourself?
  - What is your living situation? How has that changed, if at all, since classes went online?
  - Can you tell me your ethnicity and anything else about your identity that may affect your experience right now?
  - Do you consider yourself to be part of an underrepresented or disadvantaged group?
  - In what contexts are you underrepresented (e.g., as a woman in CSE but not in yoga)?
- **Health**
  - Do you have any health concerns or disabilities?
  - Is there a name or diagnosis you use? We totally respect your experience whether or not there is a diagnosis? When did that become part of your identity/when was that diagnosis made?
  - If you are registered with DRS, what accommodations do you receive? Have you negotiated those during COVID?
  - How did that impact your experience with COVID? How does that play out in terms of school and your engagement in classes?
  - Are there any technologies, tools, or strategies that you use to help you access classes and manage daily life related to your disabilities? (prompt assistive technologies)
  - How has that changed?
  - Are you using any technologies differently than you were before? Accessibility workarounds or hacks?
  - What other technologies are you using now (e.g. canvas, etc)? Are those all accessible for you?
  - We are particularly interested in how things have changed with regard to your health challenges since classes have gone online. For example, if any new challenges have arisen as you manage your classes and academic goals, if you have developed new

approaches to studying and learning, and if you are using technology differently now.
  - Have you been able to receive medical (either regular or unplanned) services? How has the pandemic changed these experiences? (e.g., are you now more or less likely to see your doctor)
  - Have you been able to receive necessary goods (e.g., groceries, medications)? Have you felt safe doing so?
  - How are you getting information and resources about health concerns? Or help resolving concerns? Is that information accessible?
  - Have you been able to talk about your health concerns with others?
  - Have you been able to get vaccinated? How has that been for you?
  - Would you have felt comfortable telling someone if you thought you were exposed to the virus?
  - What has made these conversations possible or difficult?
  - How has this changed since the start of social distancing? For example, do you think it has become easier or harder to tell someone if you are not feeling well?
- **Educational struggles**
  - Tell me about how this has all been for you academically – how classes going online, etc has been for you.
  - Have you had concerns about how you will be affected academically? How have online classes and instructional support been for you?
  - What concerns do you have about this semester? How anxious are you about this? (grades, major admission, requirements, graduation, financial aid status) If not already covered, discuss how they are using technology, how this is/isn't working.
- **Housing**
  - Have you had concerns or uncertainty about where you will be living? Any concerns about the safety of where you have been living?
- **Finances**
  - The virus has had an enormous financial toll. Can you talk about financial difficulties

that you've faced as a result of the virus, and how those play into financial challenges that you already had? Have you been able to find access to the technologies you need for education and other communication? Has your family experienced financial stress that has affected you?

  - I'd like to understand if your financial concerns are immediate (are your basic needs being met right now?), in the near future (not sure about rent in a couple of months) or in the more distant future?
  - Do you, or did you recently, have a job outside being a student? What is/was that?
  - Concerns about the future: Do you have concerns about how life might change?

- **Social Connection**
  - Relational health and example interactions
    * Can you describe some specific interactions you've had that were particularly significant/salient to you during this time? Any that were particularly good? Was there any special effort or any change in how you used technology? How did it occur to you to try this?
    * Were there any interactions in which you could comfortably express your feelings? Did you feel understood or cared for by the other person? Can you describe some of those interactions that brought a feeling of closeness? Was there any way in which you used technology to establish understanding?
    * How about not interactions that were stressful/ did not go so well?
    * Have you noticed any barriers to expressing your feelings now? Are there any ways in which it is harder or easier to do this now?
    * How have your strategies for feeling connected changed since the start of social distancing? Can you give an example?
    * Are there other celebrations you handled differently this year (e.g., birthdays)?
  - Relationships and changes in communication strategies
    * I'd like to get a sense of your relationships. Think about an inner circle of your closest contacts, a middle circle, and an outer circle of communities that you feel

part of. Starting with the first inner circle, what differences are there in how you've stayed connected? Can you comment on any changes in the amount of communication or quality of connection? How has the medium of communication changed (e.g. f2f v. online)? Please share examples of specific interactions.

    * How about the middle circle?
    * And the outer circle?
    * Please describe any specific technologies you've used and anything you've done to make them fit your needs for communicating with individuals or groups in these circles.

- Activities
    * I'd like to make sure we've included people you spend time with for different activities. I'll list some general categories and, if those are relevant to you, please describe how you used to interact with others for these activities (if at all) and how you do so now. For example, some people who used to go to yoga classes may stream those same streaming yoga classes. As above, please describe the technologies you've used and anything you've done to adapt them to your situation. For all of these, explore in relation to disability
        · Exercising
        · Going out/socializing
        · Studying
        · Interests/Hobbies
        · Faith or spiritual activities
        · Watching videos or listening to music?
        · Intimate connections: dating, flirting, hooking up, meeting people
        · Other
    * Would you like more social interaction than you have been having?
    * Typically, would you describe yourself as very social or less so?
    * Have you felt lonely? Did you feel lonely before social distancing? What changes

have you noticed? Have you felt left out?

    * Have you felt a sense of solidarity or togetherness in your community regarding covid-19?
    * Have you found ways to help and support others? Can you describe those?
    * Has anyone else reached out to you in a way that really helped you? Can you describe that?

- **Emotion Wellbeing**
  - What behaviors have been helpful for your emotional wellbeing during this time?
  - Are there any ways that you have used technology to support yourself emotionally (e.g., streaming videos for pleasant distraction, sharing articles or memes with friends to feel support, headphones to create boundaries in a household, etc). Please share specific examples.
  - Are there ways you've used technology or others have used it during this time that have jeopardized your wellbeing (e.g., constantly reading the news)
- **Biases**
  - Have you seen any changes in the way others relate to you based on ethnicity, or age, health issues or vice versa? Examples?
- **Privacy**
  - How have you sought out information about privacy (eg. related to health info)?
  - When you think about your health and telling others if you are not feeling well, do you (or have you) had privacy concerns?
  - Have you changed the way you share (e.g., loosened settings or chosen to share more in specific situations) health data and other personal data during the pandemic? Do you imagine changing that at any point?

Appendix B

Appendix: Investigating Harms in Current Behavioral Sensing Practice

B.1 Details for Section 3.3

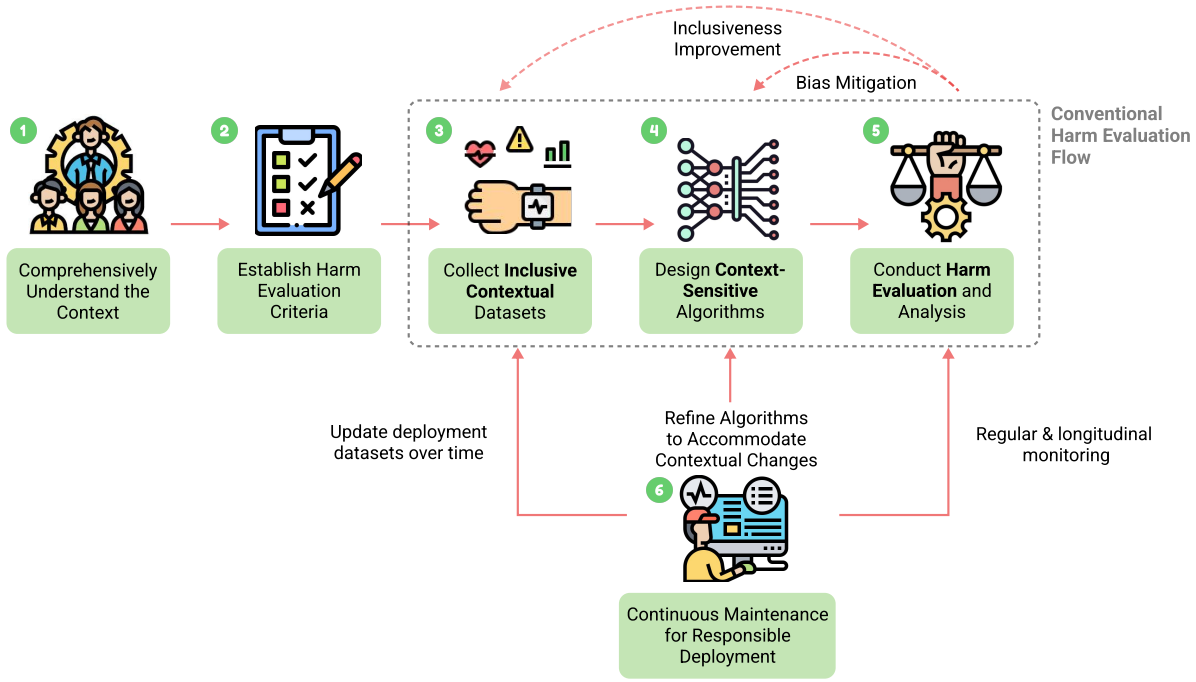


Figure B.1: Overview of the framework for evaluating and mitigating harms in behavioral sensing pipelines. Steps 3, 4, and 5 cover the conventional evaluation flow for identifying harms.

B.2 Details for Section 3.4

Table B.1: Reproduction results. Balanced accuracy of the nine depression prediction algorithms on four datasets: DS1 and DS2, which were used in prior work [Xu et al., 2023], and DS3 and DS4, which are newly reported in this study. The comparison of algorithm performance on DS1 and DS2 ensures the reliability of our fairness evaluation. The Δ column represents the difference between our reproduced results and the previously reported results.

Algorithms	DS1 (2018)			DS2 (2019)			DS3 (2020)	DS4 (2021)
	Prior results	Our results	Diff (Δ)	Prior results	Our results	Diff (Δ)	Our results	Our results
Wahle *et al.* [Wahle et al., 2016]	0.526	0.538	0.012	0.527	0.518	-0.009	0.514	0.514
Saeb *et al.* [Saeb et al., 2015a]	0.539	0.539	0.000	0.508	0.513	0.005	0.588	0.500
Farhan *et al.* [Farhan et al., 2016a]	0.552	0.552	0.000	0.609	0.609	0.000	0.563	0.609
Canzian *et al.* [Canzian and Musolesi, 2015a]	0.559	0.538	-0.021	0.516	0.516	0.000	0.541	0.502
Wang *et al.* [Wang et al., 2018a]	0.566	0.565	-0.001	0.500	0.500	0.000	0.577	0.516
Lu *et al.* [Lu et al., 2018a]	0.574	0.574	0.000	0.558	0.558	0.000	0.611	0.553
Xu *et al.* - Interpretable [Xu et al., 2019b]	0.722	0.688	-0.034	0.623	0.667	0.044	0.833	0.733
Xu *et al.* - Personalized [Xu et al., 2021]	0.723	0.753	0.030	0.699	0.690	-0.009	0.791	0.686
Chikersal *et al.* (removed) [Chikersal et al., 2021]	0.728	0.618	-0.110	0.776	0.670	-0.106	0.581	0.641

B.2.1 Evaluation Study 1: Example of the Statistical Evaluation and Experimental Implementation

Example of the Statistical Evaluation. In Figure B.2, we present an illustrative example to demonstrate our approach to fairness evaluation. In this example, we generated synthesized ground-truth labels and predictions from an algorithm for a sample of 20 individuals. Among these individuals, 6 are part of the protected group, while 14 belong to the unprotected group (as shown in Figure B.2a). We assigned a value of "1" for accurate predictions and "0" for inaccurate predictions based on the correctness of the predictions.

Figure B.2b visualizes the distribution of predictions for both the protected group (represented by "x") and the unprotected group (represented by "•"). The circle in the figure represents the distribution of predictions, with the left side indicating cases where all ground truth values are positive (representing individuals with depression in our case study), and the right side representing cases where all ground truth values are negative (representing individuals without depression in our case study). The accuracy of the predictions is indicated by the color, with green representing correct predictions by the algorithm and red representing incorrect predictions.

In this example, when considering the disparity in accuracy, the algorithm made incorrect predictions for 4 out of 6 individuals from the protected group (represented by the red "x" marks among all both red and green "x" marks). Conversely, for the unprotected group, the algorithm made incorrect predictions for 3 out of 14 individuals (depicted by the red "•" marks among all red and green "•" marks). To assess the statistical significance of these disparities, we conducted the Mann-Whitney U test in combination with the B-H correction. Specifically, we applied this test to the 2 "1" values and 4 "0" values corresponding to the protected group, as well as the 11 "1" values and 3 "0" values corresponding to the unprotected group.

When examining the difference in false negative rates, the relevant information for statistical analysis is contained in the left portion of the circle depicted in Figure B.2b. Specifically, we conducted a statistical test on the 1 "1" value and 3 "0" values in the protected group, as well as the 4 "1" values and 1 "0" value in the unprotected group. Similarly, an evaluation of the disparity in false positive rates was conducted on the marks on the right side of the circle in Figure B.2b.

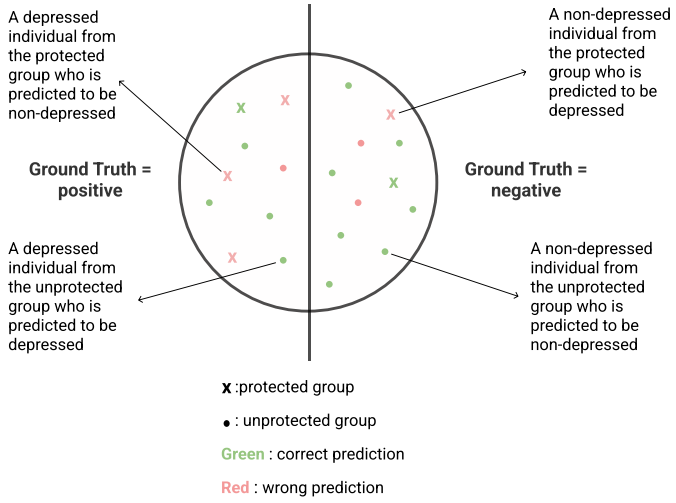
Experimental Implementation We applied the two evaluation criteria to evaluate the fairness of the eight depression detection algorithms. We provide a detailed explanation of our statistical analyses to capture disparities in accuracy, false negative rate, and false positive rate below (an example of this approach can be found in above). We provide open access to our evaluation codebase to enable reference and reproducibility for future research.

To perform the B-H correction, we first calculated the p values for all attributes using the Mann-Whitney U test. Then, we arranged the p values in ascending order and assign ranks to them, with the smallest p value receiving rank 1, the second smallest receiving rank 2, and so on. Next, we calculated the adjusted q values for each individual p value using the formula: $(i/m) \times Q$, where i is the rank of the individual p value, m is the total number of tests, and Q is the false discovery rate, 0.05. Finally, we compared the original p values to the calculated q values. Attributes with a p value smaller than the corresponding q value and less than 0.05 were considered to have significant differences.

To examine potential disparities across various groups of one algorithm, we employed a sys-

PID	True Label	Prediction	Protected Group	Assigned Value
1	1	0	Yes	0
2	1	1	Yes	1
3	1	1	No	1
4	0	0	No	1
5	0	0	No	1
6	1	1	No	1
7	1	1	No	1
8	0	1	Yes	0
9	0	0	No	1
10	1	0	Yes	0
11	0	1	No	0
12	0	0	Yes	1
13	1	0	Yes	0
14	1	1	No	1
15	0	0	No	1
16	0	0	No	1
17	1	0	No	0
18	0	0	No	1
19	0	1	No	0
20	0	0	No	1

(a) Synthetic data for 20 individuals.

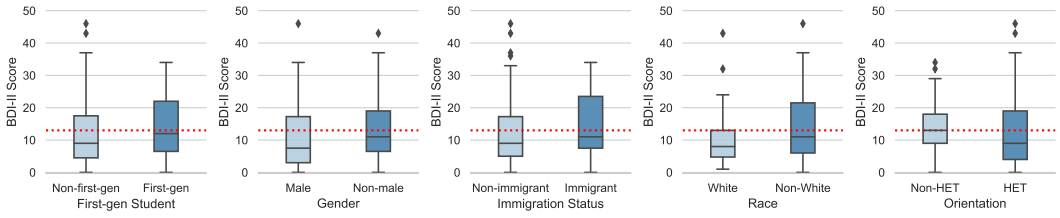


(b) Visualization of prediction distributions and disparities.

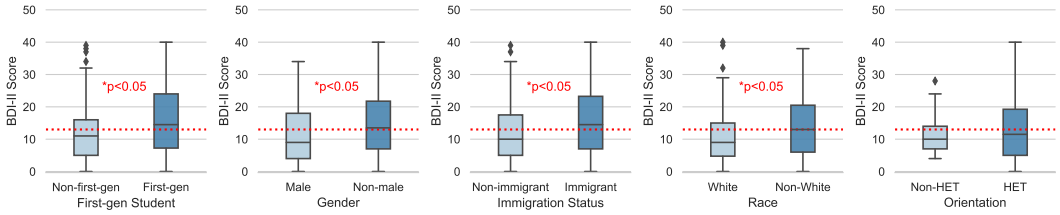
Figure B.2: Example of fairness evaluation based on the disparities in accuracy, false negative rate, and false positive rate. (a) shows the synthetic data for 20 individuals, with 6 belonging to the protected group (represented by "x" marks) and 14 belonging to the unprotected group (represented by "•" marks). (b) visualizes the distribution and disparities of predictions for both groups, where correction predictions are depicted in green and incorrect predictions in red.

tematic approach. Initially, we categorized algorithm predictions based on their correctness, assigning a value of "1" to instances where the algorithm accurately predicted the ground truth and a value of "0" to instances where the algorithm falsely predicted the ground truth. Subsequently, we applied the Mann-Whitney test in conjunction with the B-H correction to different subsets of the "0" and "1" values to evaluate the following three hypotheses. First, we conducted a thorough analysis to determine whether the algorithms exhibited comparable accuracy in predicting the ground truth for both the protected and unprotected groups, aiming to evaluate potential disparities in accuracy. To achieve this, we performed the Mann-Whitney test with the B-H correction on the complete set of "0" and "1" values. Second, our assessment focused on whether the algorithms demonstrated similar false negative rates in predicting the ground truth for both the protected and unprotected groups, to identify potential disparities in false negative rates. To accomplish this, we conducted the same test on the subset of "0" and "1" values where the ground truth labels were positive. Similarly, we proceeded to investigate whether the algo-

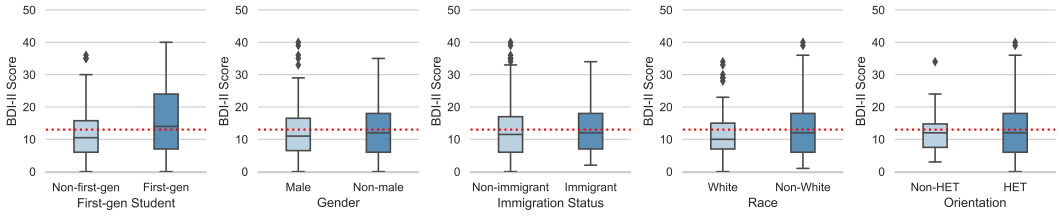
rithms displayed comparable false positive rates for both the protected and unprotected groups. This was achieved by applying the same test on the subset of "0" and "1" values where the ground truth labels were negative.



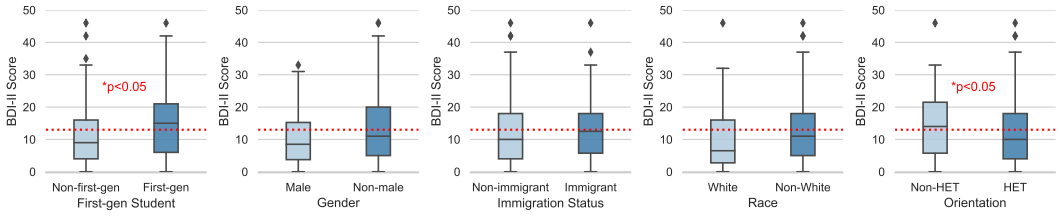
(a) DS1 (Year 2018, Pre-COVID)



(b) DS2 (Year 2019, Pre-COVID)



(c) DS3 (Year 2020, COVID Year)



(d) DS4 (Year 2021, COVID Year)

Figure B.3: Comparisons of depression (BDI-II) scores for different groups of four datasets. The red dotted line indicates the cutoff point (i.e., 13) for BDI-II scores, which is used to distinguish between students with at least mild depressive symptoms (BDI-II score >=13) and those without (BDI-II < 13). Significance levels after Benjamini-Hochberg (B-H) correction are marked with an asterisk ($*p < 0.05$) in red on the subplot. First-gen, BA, and HET represent first-generation college students, bachelor, and heterosexual, respectively.

Appendix C

Appendix: Building Human-Centered Academic Performance Prediction Models

C.1 Details for Chapter 4

C.1.1 A Review of Algorithmic Bias

There are several common definitions of subpopulation-based algorithmic fairness and corresponding evaluation metrics. We review the three most applicable fairness measures below with respect to a binary classification setting. We use all three measures to assess the algorithmic fairness of our approaches.

Demographic Parity [Barocas and Selbst, 2016]. Also commonly referred to as *Independence* and *Statistical Parity*, it requires the prediction of positive outcome, $\hat{Y} = 1$, to be the same regardless of whether the person is in a protected (e.g., female, disabled, and underrepresented minority) group ($S = 1$). Note that one disadvantage of *demographic parity* is that a fully accurate classifier may be seen as biased when the ratios of actual positive outcomes of the groups differ [Pessach and Shmueli, 2022]. Mathematically, it is computed as follows:

$$P[\hat{Y} = 1|S = 1] = P[\hat{Y} = 1|S \neq 1],$$

Equalized Odds [Zafar et al., 2017]. This measure requires the protected and unprotected groups to have the same rates for True Positives (TPs) and False Positives (FPs) [Mehrabi et al., 2021]. It was designed to overcome the disadvantage of *demographic parity* described above [Hardt et al., 2016]. Mathematically, it is computed as follows:

Table C.1: Reviewed academic performance prediction work sorted by amount of **Data** needed for prediction. The prediction **Task** is either classifying students into groups (such as below and above 3.2, in our case) or regression (continues GPA). All papers focus on end-of-term GPA except [Nghe et al., 2007], which detects end-of-year GPA and [Wang et al., 2015a], which detects cumulative GPA. The data set used as **Input** for each paper includes logs of online learning system use, student academic records, behavioral data, and self reports; the **Metrics** for assessing the model varied significantly, making comparison difficult. Half of the prior work did not consider model **Explainability**. Most of prior work did not consider model **Generalizability** for their models. No prior work considered **Fairness** of their models to marginalized student groups. **Ref.** represents for reference.

Ref.	Data	Task	Input	Metrics	Model Explainability	Model Fairness	Model Generalizability
[Nghe et al., 2007]	A year	End-of-year GPA (2-class, 3-class & 4-class)	Academic records; admissions information	Accuracy≈ 72% (4-class); 80% (3-class); 93% (2-class)	×	×	×
[Bravo-Agapito et al., 2021]	A term (14-17 wks)	End-of-Year GPA (continuous)	Learning Management System Log data; academic records; demographics	$R = 0.677$ $R^2 = 0.458$	✓	×	×
[Yao et al., 2019]	A term (14-17 wks)	Term GPA (continuous)	Campus smart card logs; academic records	Avg r=0.43, SD=0.01	✓	×	×
[Qu et al., 2019]	14 weeks	Term GPA (2-class)	Learning Management System Log Data	Accuracy=93%, Recall =0.95	×	×	×
[Sukhbaatar et al., 2019]	12 wks	Term GPA (2-class)	Learning Management System Log data; academic records	Avg accuracy≈92%, Avg sensitivity=65%, Avg precision≈75%, Avg F1=66%	×	×	×
[Wang et al., 2015a]	10 wks	Cumulative GPA (continuous)	Behavioral data from sensor; self-reports	MAE=0.18, r=0.81, R^2=0.56	✓	×	×
[Lara et al., 2014]	10 wks	Term GPA (2-class)	Online Learning Log data	Accuracy=94%, Precision=0.82, Recall=0.90, Specificity=0.95	×	×	×
[Lu et al., 2018b]	6 wks	Term GPA (continuous)	Learning Management System Log data; academic records	PMSE=159.71, R^2=0.56	✓	×	×
[Waheed et al., 2023]	5 wks	Single-class GPA (2-class)	Online Learning Log data, demographics, assessment-relatd data	Accuracy=69% Precision=0.70 Recall=0.70 AUC=0.71	✓	×	×
[Yu et al., 2018]	5 wks	Term GPA (2-class)	Academic records; self-evaluation comments	Accuracy=71%, F1=0.71	×	×	×
[Sano et al., 2015]	4 wks	Term GPA (2-class)	Behavioral data from sensor; self-reports	Accuracy=92%	✓	×	×
[Chen and Cui, 2020]	4 wks	Single-class GPA (2-class)	Learning Management System Log data	AUC=0.75 (original) AUC=0.63 (unseen data)	×	×	✓

$$P[\hat{Y} = 1|S = 1, Y = 0] = P[\hat{Y} = 1|S \neq 1, Y = 0], P[\hat{Y} = 1|S = 1, Y = 1] = P[\hat{Y} = 1|S \neq 1, Y = 1],$$

Equal Opportunity [Hardt et al., 2016]. This is also commonly referred to as *recall* or *sensitivity*. It is less strict than *equalized odds*, which only requires the protected and unprotected groups to have equal true positive rates (or false negative rates). Mathematically, it is computed as follows:

$$P[\hat{Y} = 1|S = 1, Y = 1] = P[\hat{Y} = 1|S \neq 1, Y = 1].$$

To set fairness criteria, the definition of disparity that is expressed as a difference is often considered [Ahmad et al., 2020; Kobayashi and Nakao, 2021]. For example, the *demographic parity difference* is defined as the difference in the probability of prediction between the two groups. Similarly, one may calculate an *equalized odds difference*, or the greater of two metrics, True Positive Rate (TPR) difference and False Positive Rate (FPR) difference between the two groups; and an *equal opportunity difference*, which only compares the TPs between the unprotected and protected groups. In each case, a difference of 0 indicates that the model is perfectly fair to the protected trait (it favors neither the protected nor the unprotected group).

Another common fairness criteria is to compute a ratio between groups [Ahmad et al., 2020]. For example, the *demographic parity ratio* (also called disparate impact [US Equal Employment Opportunity Commission et al., 1978]) is defined as the ratio between the probability of positive prediction for the unprotected group and the probability of positive prediction for the protected group. A ratio of 1 indicates that the model is fair relative to the protected trait (it favors neither the protected nor the unprotected group). In US law, a value of demographic parity ratio (or disparate impact) more than 0.8 indicates that there is not an unfair situation (80% rule) [US Equal Employment Opportunity Commission et al., 1978; Kabacoff et al., 1997]. Similarly, *Equalized odds ratio* is defined as the smaller value of two metrics, TPR ratio and False Negative Rate (FNR) ratio [FairLearn Contributors, 2022], where TPR and FNR ratios are calculated as the rate of the unprotected group divided by the rate of the protected group. An equalized odds ratio of 1 means that all groups have the same true positive, true negative, false positive, and false negative rates, respectively [FairLearn Contributors, 2022]. *Equal opportunity ratio* is calculated as the ratio of TPs between the unprotected and protected groups. A value

of 1 means that all groups have the same TPR, and that the model is within the "fair" range relative to the protected trait.

Table C.2: Passive-sensing data and extracted low-level behavioral features.

Source	Sensor	Sampling frequency	Low-level Behavior features
Smartphone	Physical activity	Every minute	Most common activity, number of activities
	Application usage	Event-based	Number of used apps, most commonly used app, most common app category, apps used per minute
	Battery	Event-based	Number of charging sessions, total charging time
	Bluetooth	Every 3 minutes	Number of scans, number of unique devices, number of scans of least frequent device, number of scans of most frequent device, etc.
	Calls	Event-based	Number of incoming calls, number of outgoing calls, number of missed calls, duration of incoming calls, duration of outgoing calls, etc.
	Locations	Every minute	Total distance traveled, time spent at/near home, average traveling speed, percentage of time moving, time spent at top 3 location clusters, etc.
	Location map	Every minute	Time at exercise-labeled places, time at food-labeled places, time at fraternity-labeled places, time at greens-labeled places, time at living-labeled places, time at study-labeled places, etc.
	Screen	Event-based	Sum duration of phone interactions, average duration of phone interactions, standard deviation of interaction durations, time of first unlock event, time of last unlock event, number of unlocks per minute, etc.
	WiFi	Every 3 minutes	Number of unique access points, most frequent access point
Fitbit	Sleep	Every minute	Time in bed, awake duration, asleep duration, restless duration, sleep efficiency, etc.
	Step count	Every minute	Total step count, number of active bouts, average duration of active bouts, average steps per active bout, start time of longest active bout, etc.

C.1.2 Implementation of High-Level Behavior Features

Activity Duration. We implemented this feature by grouping consecutive activity data samples with non-stationary labels (i.e., on foot, walking, running, and on bicycle) into activity bouts, and then computing the total activity duration of the student by summing up the duration of the bouts .

Study duration and study focus. We included any dwelling time of 20 minutes or greater at study labeled locations (e.g., libraries, teaching buildings, and cafes) in the estimation of a student's study duration. We considered students being stationary at the study locations to be

more focused on studying. By fusing location data and activity data, we calculated study focus as the percentage of the dwelling time with stationary activity labels (i.e., still and tilting) with respect to the total study duration.

Dorm duration. We computed this feature as the total amount of time a student spend at places labeled as "dorm" or "living".

Party duration. We considered students staying at the fraternity houses on campus any time from 6pm to 12pm the next day with a dwelling time of 30 minutes or above to be partying and calculate this feature by summing up the dwelling time. We excluded the students who live at the fraternity houses from the calculation.

Indoor and outdoor mobility. Similar to [Wang et al., 2015a], we fused location and activity data and calculate indoor mobility as the total amount of time when a student is walking or running indoors. We calculated outdoor mobility as the total distance traveled by the student when he/she is outdoors.

Class Attendance. We computed class attendance related features using both students' class schedules and location data. Similar to location map features, we hand labeled the locations of all teaching buildings on campus. For each class period a student was scheduled to attend, we compared the student's location during the class time against the teaching building of the scheduled class. We calculated the amount of time a student was at the correct building as a percentage of the total class duration, and considered the student attending the class only if the percentage is more than 50%.

Behavioral Change. We divided each academic term into individual weeks and capture a student's overall behavioral changes within each week. We followed a similar approach to [Wang et al., 2015a] and computed slopes and breakpoints on a weekly basis for all the above-mentioned behavior features. We defined Thursday as the midpoint of each week (starting on Monday), and fit linear regression models to the data of the first half (midpoint excluded), second half (midpoint included) and the entire week, respectively. We designated the slopes of the above three linear regression models as first-half slope, second-half slope, and slope all. Note that, slope captures 1) the direction of behavioral change (i.e., increases or decreases in sleep duration) and

2) magnitude of the behavioral change (i.e., steep or gradual changes in sleep duration) within the first week, as well as the first half (Monday to Wednesday) and second half (Thursday to Sunday) of the week. Separate from the midpoint, we also computed breakpoints that capture the specific day in the first week when a student's behavioral pattern shows a directional change (i.e., the day when their sleep duration increases or decreases).

C.1.3 Data Preprocessing

Common Data Cleaning. Before modeling, we assigned each participant a unique participant ID to ensure privacy, with all analyses conducted using anonymized data. Missing values primarily arose due to data collection challenges, such as app crashes, phones running out of battery, or participants failing to comply with study protocols, like not wearing their Fitbit or skipping questionnaires. Features that were 100% missing were removed from the dataset. We handled numeric outliers by capping them based on the interquartile range (IQR) calculated for each student individually. Categorical features were transformed using one-hot encoding to prepare the data for model training.

Customized Data Preprocessing for The LR Approach. ***Missing Value Handling.*** During data preprocessing, features with 100% missingness across the dataset were removed. For remaining features with missing values, we tested two imputation methods: (1) imputing missing values with a default value (999), and (2) imputing values with the mean of the training set. Based on model performance in 2018, we selected the second method. During leave-one-subject-out cross-validation (LOSO-CV), if a feature in the training set was entirely missing, the default value of 999 was used.

Class Imbalance Handling. Our data from 2018 and 2019 is imbalanced, with only 23% and 32% of students having lower GPAs, respectively. To address this, we experimented with SMOTE and ADASYN for oversampling the minority class in the training set to balance the classes. SMOTE was chosen based on the 2018 model performance [Chawla et al., 2002].

Collinearity Removal. To avoid issues of collinearity that could distort model estimation, we removed features from the training and test sets that were highly correlated ($|r| > 0.7$) based on

training set data [Dormann et al., 2013].

Feature Selection. We employed correlation-based feature selection (CFS [Hall, 1999]) to identify features significantly correlated with end-of-term GPA ($p < 0.05$). For each round of LOSO-CV, we performed a grid search to determine an optimal correlation threshold r, selecting the r value that maximized the performance advantage ($a_{diff} = a_{test} - a_{train}$). We note that while the use of test data in determining r introduces some leakage, this was only during feature inclusion, and no leakage occurred when applied to the 2019 data.

Customized Data Preprocessing for The 1D-CNN Approach. ***Missing Value Handling.*** Since the data used in the deep learning model is time series data, we employed forward filling to impute missing values initially, followed by backward filling for any remaining gaps. This is a standard technique for handling missing data in time series [Che et al., 2018]. Unlike the LR pipeline, which used mean imputation, this approach ensures that the imputed values reflect the temporal sequence of the data, as aggregating by week (as done in LR) does not apply to continuous time series data.

Class Imbalance Handling. To address class imbalance, we adopted a simple oversampling approach by randomly duplicating samples from the minority class (i.e., low performers) to equalize the ratio between the two classes (1:1) in the training set.

Data Standardization and Transformation. We standardized all features and transformed the data into a three-dimensional time series format, suitable for deep learning models, structured as [number of participants, number of days, number of features].

Architecture of 1D-CNN Model. The architecture of the 1D-CNN model includes a single 1D convolutional layer (1D-CNN) followed by a rectified linear unit (ReLU) activation function. To prevent overfitting, we applied a dropout layer immediately after the 1D-CNN layer, masking 85% of its output [Gal and Ghahramani, 2016]. This is followed by a max pooling layer, which reduces the spatial size by applying a max filter to non-overlapping subregions of the dropout layer's output. The pooled output is then flattened into a single vector via a flattening layer. Finally, the model contains two dense (fully connected) layers: the first uses a ReLU activation

function, while the output dense layer employs a softmax function to return class probabilities for the binary classification task. The model was optimized using the Adam optimizer, with categorical cross-entropy as the loss function. The training process used 150 epochs with a batch size of 6. Hyperparameters, including a learning rate of 0.0001, were selected using grid search. Additionally, early stopping was employed, halting training after 10 steps without improvement.

C.1.4 Academic-Related Patterns and Factors

Below, we summarize behavioral patterns and factors and discuss their implications for early intervention strategies. These results are derived from Tables 4.3 and 4.4, where features suggesting similar patterns have been grouped together.

Weekday vs. Weekend Behavior.s One interesting observation is that many of the identified behavioral shifts (breakpoints in daily routines) during the first week of the Spring term, for both years, occur on Thursdays (e.g., 2018-R7 to 2018-R10, 2018-R18, 2019-R2, 2019-R3). This suggests that students' weekend behaviors may begin on Fridays rather than Saturdays for a substantial portion of the population. This distinction could offer valuable insights for targeted interventions, as shifts in behavioral patterns earlier in the week may indicate opportunities for academic support or engagement efforts before the weekend.

Class Attendance. Not surprisingly, average class attendance during week one is positively associated with end-of-term GPA (2018-R14). This finding aligns with existing literature, which shows a strong relationship between class attendance and both individual course grades and overall GPA [Credé et al., 2010; Boumi and Vela, 2021]. This consistency reinforces that the features identified in our study as predictors of academic performance are meaningful and worth exploring further. It also suggests that educators should take note of students' attendance early in the term and proactively check in with those who are not attending to understand potential barriers. Since flexibility in attendance is critical for addressing accessibility needs [Zhang et al., 2022], such outreach should avoid mandating physical presence, as this could place additional stress on students with disabilities or those experiencing mental health challenges.

Phone Usage. An increase in phone usage is negatively associated with students' end-of-term GPA (2018-R7 and 2019-R19), a finding supported by prior research showing a negative correlation between smartphone usage and academic performance [Lepp et al., 2015; Giunchiglia et al., 2018; Yeboah and Ewur, 2014]. Our results suggest that this effect is particularly pronounced on *weekdays*, adding nuance to the general understanding of this relationship. Studies show that in-class phone use can significantly hinder student performance [Sumuer, 2021], with in-class usage being nearly double that of outside-classroom use [Felisoni and Godoi, 2018]. Whether phone use is a cause or consequence of struggling academically—or perhaps a related factor such as stress—is unclear, but these patterns suggest that students who are distracted by their phones during weekdays may not be setting themselves up for academic success.

Phone usage is also linked to stress [Višnjić et al., 2018], and excessive use can act as a negative coping strategy [Augner and Hacker, 2012], which is further supported by our finding that feeling helpless in difficult situations (2018-R12) is negatively associated with GPA. This highlights an opportunity for student wellness programs or new student orientation initiatives to address the role of stress, coping strategies, and phone use in academic success. Interestingly, our findings also suggest that phone usage during weekends may help students unwind, as more frequent use in the evenings and nights later in the weekend is positively associated with academic outcomes (2018-R1 and 2018-R15), indicating that weekend phone use might serve as a way to relax after a week of hard work.

Time Spent at Different Locations. An increase in time spent at living places, such as home or dorms, during evenings or at night on weekdays is positively associated with end-of-term GPA (2019-R3 and 2019-R16). This may reflect the value of students engaging in campus life during the day and then spending time with roommates or dorm mates in the evenings, possibly studying or socializing. This type of evening engagement could alleviate feelings of isolation or a lack of belonging, as students who reported knowing less about school than their peers (2018-R13) exhibited a negative association with GPA. However, longer durations spent at living places during the day or in the morning on both weekdays and weekends were associated

with lower end-of-term GPA (2018-R9, 2018-R25, 2018-R29, 2019-R6, 2019-R11, and 2019-R21). This suggests that while evening time at living places may be beneficial for academic success, excessive time spent indoors during the day or morning, especially on weekends, could detract from opportunities to engage with the broader campus environment, which might support students' academic and social integration.

Additionally, an increase in time spent at exercise locations throughout the day on weekdays is positively associated with academic performance (2018-R22), suggesting that physical activity during the week may contribute to better academic outcomes. However, spending more time at exercise places during the evening on weekends is associated with worse end-of-term GPA (2019-R4 and 2019-R5), which may indicate a potential disruption to academic focus or preparation for the upcoming week. Similarly, time spent in green spaces during the afternoon and evening is positively associated with academic performance (2019-R2 and 2019-R17). This could reflect the value of engaging with the campus environment, benefiting from outdoor activity and social interaction, especially during times that support mental well-being without conflicting with academic obligations like class attendance. These patterns highlight the importance of balanced engagement in physical and outdoor activities during the week, while suggesting that weekend activities might need to be managed to avoid negatively affecting academic outcomes.

Furthermore, an increase in time spent at food places in the evening on weekends is positively associated with student academic performance (2019-R12 and 2019-R24), suggesting that social or leisurely activities in such settings may serve as a beneficial break for students. Interestingly, while party duration at night during the first week is negatively associated with academic performance (2019-R22), time spent at Greek houses on weekend nights—regardless of whether students live in them—is positively associated with end-of-term GPA (2019-R15). Although previous research has suggested that Greek membership may negatively impact academic performance [Grubb, 2006], our findings indicate that moderate socializing at these locations may not be inherently harmful. This suggests that relaxing at a party or gathering during weekends can be a healthy activity for students. Further research could explore whether students who are confident in their academic performance are more likely to attend social events and examine

what types of social behaviors, including those in Greek life, are most supportive for students dealing with academic or personal stressors.

Sleep. Longer periods of restless sleep are negatively associated with student academic performance (2018-R18), as are extended periods of being awake, such as getting up early and staying awake late or pulling all-nighters (2019-R18). This is consistent with previous research that highlights the detrimental impact of poor sleep on academic outcomes [Okano et al., 2019; Gomes et al., 2011]. Interestingly, the longer the shortest duration of staying awake in a 24-hour period (2018-R6)—which could indicate a nap or a period of insomnia—is positively associated with higher end-of-term GPA. This finding warrants further investigation, as it contrasts with existing literature that shows sufficient sleep, good sleep quality, and greater sleep consistency are positively linked to academic performance [Gomes et al., 2011]. Understanding the nuanced relationship between short wakefulness periods and academic outcomes could provide deeper insights into student sleep patterns and their impact on academic success.

In addition to the above behavioral patterns that are relatively easy to interpret, we note that some location patterns are harder to explain. For example, an increase in time spent at the third-ranked location cluster (2018-R11), time spent at second-ranked location cluster (2018-R28, 2018-R30, 2019-R7, 2019-R8, and 2019-R30) at any time in a day was negatively associated with end-of-term GPA.

Self-reported Stressors. We also observed several serious stressors from students' self-reports that were, unsurprisingly, strongly negatively associated with GPA. These included issues with romantic partners (2018-R13), health concerns (2018-R21, 2019-R25), and traumatic experiences (2019-R29). Additionally, academic-related stressors such as receiving lower grades than expected in a prior term (2018-R3) and obtaining a lower GPA than anticipated (2018-R16) were also negatively associated with end-of-term GPA. These findings align with prior research that highlights the negative impact of stressors on academic outcomes [De Luca et al., 2016; Pereira et al., 2018]. The strong association between stressors and academic performance suggests that early intervention strategies focusing on mental health support and academic counseling could

be highly beneficial. Proactively addressing these issues at the beginning of the term might help students better manage their stress and improve their academic resilience.

Interestingly, we find that type of phone service provider is also associated with students' academic performance. To better understand this feature, we compare the proportion of each service provider use between students with higher and lower GPAs. We found that high performers used AT&T, Cricket, and Sprint more, while low performers were prone to use Virgin Mobile and other providers. A chi-square test of independence was performed to examine the relation between high/low performers and the service provider they used. The relation between these variables was significant, $\chi^2(6,\ N = 188) = 17.1,\ p < .01$. This certainly means that there is some other variable that is not being captured in our feature set that connects phone service provider and performance, perhaps something related to income or childhood home locale or other context, and highlights the importance of comparing features to behavioral science knowledge.