Can Data Augmentation Improve Daily Mood Prediction from Wearable Data? An Empirical Study

Neha Manjunath*
Cornell Tech
New York, United States
nm468@cornell.edu

Ze Yuan Li*
Cornell Tech
New York, United States
zl344@cornell.edu

Eunsol Soul Choi*
Cornell Tech
New York, United States
ec897@cornell.edu

Srijan Sen
University of Michigan
Ann Arbor, United States
srijan@med.umich.edu

Fei Wang
Weill Cornell Medicine
New York, United States
few2001@med.cornell.edu

Daniel A. Adler
Cornell Tech
New York, United States
daa243@cornell.edu

ABSTRACT

Mobile sensing data, approximating human behavior and physiology, can be processed by machine learning models to predict mental health symptoms. While these models are accurate in smaller samples, their generalization accuracy decreases in larger samples, potentially because it is difficult to collect enough mobile sensing and mental health outcomes data at scale to enable generalization. In this study, we hypothesized that augmenting training data with synthetic data samples could improve the generalizability of these machine learning models. We created a data augmentation system that generated synthetic mobile sensing and mental health outcomes data, and evaluated the utility of this system via the downstream machine learning task of predicting daily mood from wearable sensing data. We experimented with both simple (e.g. noise addition) and novel generative data augmentation methods, based upon conditional generative adversarial networks and multi-task learning. Our initial findings suggest that the data augmentation system generated realistic synthetic data, but did not improve mood prediction. We propose future work to validate our findings and test other methods to improve the generalizability of mental health symptom prediction models.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; • **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Mobile Sensing; Mental Health; Deep Generative Models; Wearables; mHealth

*These authors contributed equally to this research.

UbiComp/ISWC '23 Adjunct , October 8–12, 2023, Cancun, Quintana Roo, Mexico

ACM ISBN 979-8-4007-0200-6/23/10…$15.00
https://doi.org/10.1145/3594739.3612876

ACM Reference Format:
Neha Manjunath, Ze Yuan Li, Eunsol Soul Choi, Srijan Sen, Fei Wang, and Daniel A. Adler. 2023. Can Data Augmentation Improve Daily Mood Prediction from Wearable Data? An Empirical Study. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct), October 8–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3594739.3612876

1 INTRODUCTION

Behavioral and physiological data collected passively from consumer mobile technologies offers a unique opportunity to gather contextual data about patients to inform mental health symptom monitoring and treatment [22]. This opportunity has prompted researchers to develop machine learning models that process mobile sensing data to predict symptoms of mental illness [1, 15, 24]. This prior work demonstrates the feasibility of using mobile sensing data for remote mental health assessment, providing low-burden methods to passively monitor mental health and identify individuals in need of care.

Despite this potential, it is difficult to collect mobile sensing data and mental health outcomes at scale in any single study [4], potentially contributing to the poor generalization accuracy of these machine learning models – around 60%, at best – in larger cohorts [21, 23, 33]. Prior work from other areas of machine learning [27, 30] shows that *data augmentation methods* can improve model generalizability without further data collection by generating synthetic, but realistic training samples where synthetic data injection can be added to simulate a larger dataset. To this end, we developed a data augmentation system for generating synthetic mobile sensing and mental health outcomes data. We then evaluated if this system improved the performance of machine learning models that predicted mental health symptoms using mobile sensing data. Within the augmentation system, we experimented with using common, simple data augmentation techniques (eg, noise addition), as well as more sophisticated generative modeling techniques, including a novel generative model architecture, designed specifically to model mobile sensing and mental health outcomes. Our initial findings suggest that the novel generative model was able to create realistic synthetic samples, but that the data augmentation system overall, using both simple and generative methods, was not able to improve mental health prediction models. From these experiments,

we discuss opportunities for future work to further evaluate our findings and improve mental health symptom prediction model generalizability.

2 RELATED WORK

Previous studies have explored various data augmentation methods, and many of these methods have been applied to predict health outcomes in longitudinal data, similar to mobile sensing data. A recent survey categorized these techniques into simple methods that directly transform data through actions like deleting, warping, or adding random noise, as well as more advanced statistical or deep learning based approaches that involve training a model to learn the underlying data distribution and generating synthetic data based upon this model [32]. These methods often draw inspiration from data augmentation techniques used for more-common machine learning tasks like image recognition [27]. For instance, [30] utilized techniques such as jittering (adding random noise) and scaling (multiplying by random noise) to improve the performance of a machine learning model that monitored Parkinson's Disease symptoms.

Recent work has experimented with more sophisticated methods, based upon generative models, for augmenting mobile sensing data. For example, Li et al. proposed a generative adversarial network (GAN) based model, called ActivityGAN, to generate synthetic accelerometer data [18]. The authors assessed the utility of the synthetic data within various human activity recognition tasks. Similarly, Haradel et al. generated synthetic ECG and EEG data using a GAN, and showed how the augmented data improved the performance of machine learning models developed to detect heart attacks and seizures [13]. While GANs may accurately generate synthetic accelerometer, ECG, or EEG data, they face challenges in generating multivariate, multimodal behavioral data due to "mode collapse', where synthetic data generated by GANs are only represent a single distribution mode [2]. This calls for novel methods to more-accurately capture the multimodal nature of mobile sensing and mental health outcomes data for realistic data generation.

To the best of our knowledge, only Yu and Sano have experimented with data augmentation methods to improve the performance of machine learning models that predict mental health symptoms using mobile sensing data [34]. The authors experimented with using simple, noise addition methods to augment wearable, smartphone, and ECG data, and then used this data to predict two self-reported mental health symptoms, observing only minimal improvements in model performance. Expanding upon this work, we aimed to evaluate if both simple and more novel generative model based data augmentation methods could improve mental health prediction model accuracy. To do this, we proposed a novel GAN architecture, specifically designed to augment multimodal behavioral and mental health outcomes data.

3 DATASET

In this work, we evaluated the data augmentation system within the example task of predicting the daily mood of medical interns. A medical internship is the first year of a U.S.-based resident physician training program, and is known to be stressful: residents take care of extremely sick patients, and are often required to work 24 hour shifts [5]. This accumulated stress has been linked to higher rates of depression among residents [12, 20]. Developing low-burden methods to help residents identify mental health symptoms may provide motivation to seek care. This motivates developing machine learning models to predict residents' mental health, enabling timely intervention, treatment, and ideally prophylaxis of mental illness.

3.1 Data Collection Overview

Specifically, we developed models using data collected during the Intern Health Study. The Intern Health Study is a multi-site prospective study to examine relationships among behavior, mental health and well-being during a medical internship [25, 26]. Interns employed at participating residency programs throughout the United State were able to enroll in the study online. Participants who provided informed consent were mailed a Fitbit Charge 2 [9], which continuously collected minute-by-minute heart rate, sleep, and step count data. In addition, participants installed a study smartphone application to self-report daily mood. Data collection took place over a period of 13 – 14 months, from the two months prior to the start of the internship through the year-long internship. All study procedures were approved by the University of Michigan Institutional Review Board (IRB). Collected data was exclusively used for research purposes, and participants received a Fitbit device and up to US $125 as compensation for participation.

3.2 Wearable Sensing Features

We calculated eleven daily wearable sensing features described in prior work [16, 29] from collected Fitbit data. Specifically, we approximated the daily step count, as well as the step count during the most active 10 hours of movement (M10, estimated by identifying the consecutive 10 hours each day with the most steps), and during the least active 5 hours of movement (L5). We also calculated the intraday variability, by dividing the squared difference in step count during subsequent hours by the overall variance in step count throughout an entire day [29]. High variability, M10, and low L5 signify daily rest-activity fluctuations that are important for mental health [29]. From the heart rate data, we calculated the daily average resting heart rate. We then approximated the heart rate variability (HRV), a measure of autonomic nervous response to stress [6], by inverting the heart beats per minute, giving the average time between beats, and calculating the deviation in these values. Future work can identify better methods to approximate HRV when it is not directly measured from devices. Finally, for sleep, we calculated the time spent asleep, in bed, in deep, light or REM sleep. Sleep categories (e.g. REM) are automatically estimated by the Fitbit API.

3.3 Daily Mood Prediction Task

We evaluated the data augmentation system by predicting participants' self-reported daily mood from the created wearable sensing features. Mood is an important indicator of underlying depression symptoms [17], and self-reported mood has been used as a prediction outcome in prior research using mobile sensing data to predict mental health [16, 28].

In this study, a daily mood survey was delivered to participants between 5-10PM. The survey asked participants to rate their average mood over that day from 1 (low mood) to 10 (high mood). Similar to prior work [34], we binarized self-reported mood to create a classification task. Specifically, mood scores ≥ 8 were labelled as the negative class and scores < 8 were labelled as the positive class. We used 8 as a threshold because it allowed us to create a balanced classification problem (50% positive, 50% negative samples). Future work can evaluate the effects of data augmentation in imbalanced settings, which would make it more difficult to predict specific outcomes.

4 DATA AUGMENTATION SYSTEM

In this section, we describe the data augmentation system we developed to improve the performance of mood prediction models. An overview of this system is presented in Figure 1. The system worked by augmenting training data within a personalized cross-validation procedure from prior work [1, 31]. We used this procedure because the generalization of similar machine learning models has been poor without personalizing to individual participants [21, 23, 33]. In this approach, each participant's data was temporally split such that the first x% of collected data was used to train data augmentation and mood prediction models, and the remaining (100 - x)% was used as validation data to report performance metrics. We experimented with x = 20, 40, 60, and 80. Despite the dynamic nature of this approach, the data points used for training and validation within each split (x) remained consistent across all augmentation methods. This ensured that different augmentation techniques could be compared within each split, but different augmentation techniques should not be compared across different splits. In addition, for every augmentation method, we generated synthetic data points to match the training dataset size (ratio of real to synthetic training data is 1:1).

We tested multiple simple data augmentation methods from prior work, as well as more novel data augmentation models based upon generative adversarial networks (GAN). Augmented (combined synthetic and actual training) data trained downstream mood prediction models. The mood prediction models were extreme gradient boosting (XGBoost) models, a more regularized form of gradient boosting decision trees [8]. This model class has been used in prior work to predict mental health from mobile sensing data [3, 19, 31].

4.1 Simple Augmentation Methods

Simple data augmentation methods from [14, 34] were applied to mobile sensing features, holding the mood outcome data constant. In these methods, $x \in \mathbb{R}^m$ is a standardized feature vector (i.e. $\mu = 0, \sigma = 1$) and m is the number of features. An augmented sample is described as $x \in \mathbb{R}^m$.

4.1.1 Jittering. Jittering adds random $\epsilon \sim N(0, \sigma), \epsilon \in \mathbb{R}^m$ to the mobile sensing features (see equation 1). We experimented with $\sigma \in (0.001, 0.01, 0.1, 1.0)$.

$$x' = x + \epsilon \tag{1}$$

4.1.2 Scaling. Scaling (equation 2) changes the magnitude of all the data in the dataset by a constant factor α. We experimented with multiple $\alpha \in (0.75, 0.9, 1.1, 1.25)$.

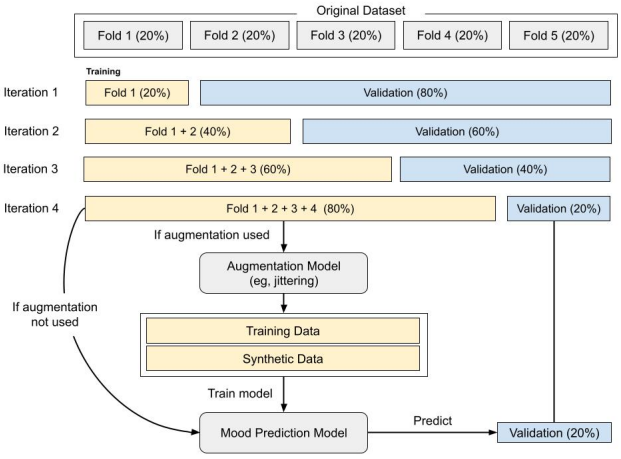

Figure 1: An overview of the data augmentation system and cross-validation procedure used in this work. We used the personalization cross-validation procedure suggested in prior work [1, 31], temporally splitting samples across individuals. For example, in "Iteration 4" the first 80% of collected data from each participant trained augmentation and mood prediction models, and models were validated using the remaining 20%.

$$x' = \alpha * x \tag{2}$$

4.1.3 Random Scaling. Random scaling multiplies each sample (equation 2) by a random $\alpha \in \mathbb{R}^m$ sampled from a Gaussian distribution $\alpha \sim N(1, \sigma)$. We experimented with $\sigma \in (0.01, 0.05, 0.1, 0.5)$.

4.2 GAN-based Augmentation Methods

We explored using GANs for data augmentation because of their demonstrated ability to generate realistic sensor data [13, 18]. We experimented with both a traditional GAN (Figure 2a), and a novel conditional GAN (CGAN) architecture developed in this work to generate multimodal, multivariate wearable sensing and mood outcomes data (Figure 2b).

4.2.1 GAN. We created a basic generative adversarial network using methods from [11]. Random noise z was sampled from a Gaussian distribution, $z \sim N(0, 1)$. The random noise was input into a fully connected neural network, G, to generate a multivariate synthetic data point $x' = G(z), x' \in \mathbb{R}^m$. This multivariate synthetic data point modeled both features and raw mood data. A discriminator fully connected neural network, D, was trained to classify whether a data point was real (x) or synthetic (x'). During model training, the generator and discriminator networks "compete". The discriminator learns to differentiate real and synthetic data, and subsequently, the generator learns to generate more realistic synthetic data points to fool the discriminator. In our study, both the generator (12, 32, 64 nodes) and discriminator (64, 32, 1 nodes) were modeled using three layered feedforward neural networks.

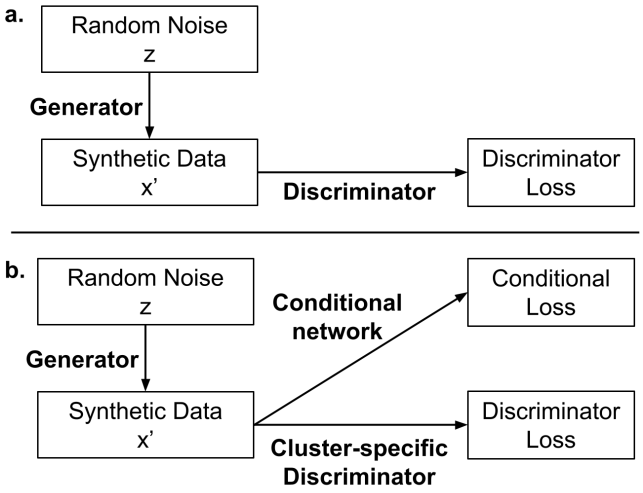

Figure 2: The (a) generative adversarial network (GAN) and (b) novel multi-task, conditional GAN model created for the data augmentation system. z is random noise drawn from a Gaussian distribution $z \sim N(0, 1)$. x' is a multivariate synthetic mobile sensing and mood data point output by the generator network.

4.2.2 Multi-task CGAN. Prior work shows that GAN models experience "mode collapse" in multimodal, multivariate behavioral data [2], only able to accurately model data around a single distribution mode. To counter this, we developed a novel conditional GAN (CGAN) architecture using multi-task learning [7], modeling different distribution modes as GAN prediction tasks.

First, we identified the number of tasks (modes) by clustering training data using K-means, varying the number of clusters, k, from 1 to 10. The optimal number of tasks (clusters) was chosen by analyzing the average within-cluster sum of squared distances between the cluster centers and each data point. We then developed a GAN model with a single generator for the entire dataset, but separate discriminators per task. In addition, the GAN had an additional conditional network and loss. The conditional network and loss was a multiclass prediction problem, predicting the specific cluster a generated data point belonged to based upon the cluster-specific discriminator. The generator was a four-layer feedforward neural network (12, 256, 512, and 1024 nodes), and each task-specific discriminator was a five-layer feedforward neural network (1024, 1024, 512, 256, 1 nodes). The conditional network was a three layered feedforward neural network (1024, 64, final node = number of clusters).

5 RESULTS

5.1 Data Overview

71,789 samples (daily wearable sensing and mood outcomes data) were collected from 1,206 participants. The median number of samples collected per participant was 43, with an interquartile range (IQR) of 13 to 94. 50.4% of samples were binarized as low mood (self-report <8). More information on the dataset and participants can be found in Table 1.

Table 1: Dataset and Participant Overview

Method	Value
Data	
Number of samples	71,789
Number of participants	1,206
Samples per participant, Median (IQR)	43 (13 to 94)
% Samples Positive Mood Class (self-report < 8)	50.4%
Demographics	
Age, Median (IQR)	27 (26 to 28)
Female	56.0%
Male	43.6%
Unknown Sex	0.4%
White	60.8%
Asian	20.6%
Multi-racial	8.3%
Black/African American	4.3%
Latino/Hispanic	3.5%
Other	2.5%

5.2 Validating the GAN Results

We first validated that the GAN and more novel multi-task CGAN generated realistic synthetic data samples by visually inspecting real and generated data, a common practice to assess GAN model performance [2, 10]. Histograms of four example features can be found in Figure 3. We found across all features that the GAN model did not generate realistic data, demonstrating signs of "mode collapse", only generating samples around the mode of the distribution. The multi-task CGAN appeared to generate more-realistic synthetic data, indicating that multi-tasking and adding a conditional loss may aid in improving wearable and mood data generation.

5.3 Mood Prediction Results

We evaluated downstream mood classification performance by calculating the area under the receiver operating curve (AUROC, see Table 2). The AUROC for baseline models without any augmentation ranged from 0.5715 with 20% training data to 0.5934 with 80%. Simple data augmentation methods, using jittering, scaling, and random scaling did not drastically baseline model performance. The GAN resulted in a lower AUROC that increased from 0.5032 with 20% training data to 0.5638 at 80%. The multi-task CGAN showed a small improvement over the GAN, with AUROC increasing from 0.5233 with 20% training data to 0.5812 at 80%.

6 DISCUSSION

In this work, we experimented with using simple and novel generative model based data augmentation methods to improve the performance of a machine learning model that classified mood using wearable sensing data. We found through visual inspection that our novel generative model, the multi-task CGAN, successfully generated realistic synthetic wearable sensing and mood data (Figure

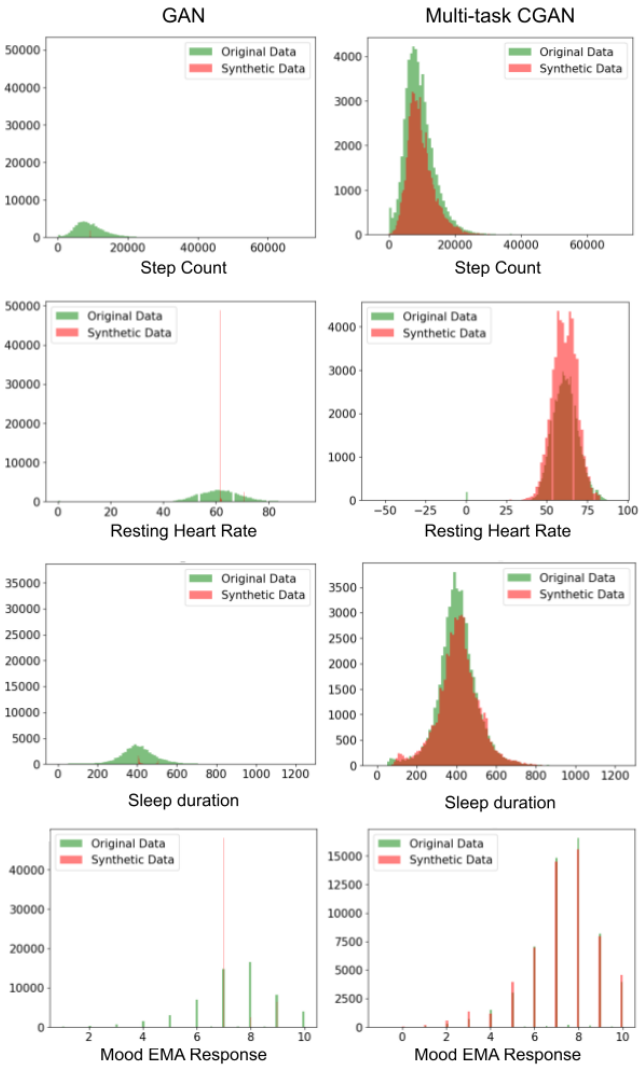

Figure 3: Example histograms for select features (x-axis label) comparing the original to the synthetic data distributions for the GAN and novel multi-task conditional GAN (CGAN).

3). Despite successfully recreating data, we found that both simple (jittering, scaling and random scaling) and generative methods did not improve mood classification (Table 2). In this discussion, we contextualize these findings with literature, and discuss opportunities for future work to further validate these results and improve model generalizability.

Our results showed that simple and deep learning based augmentation methods did not improve daily mood classification. This extends findings from prior work showing that simple data augmentation methods do not improve machine learning models predicting mental health from mobile sensing data [34]. In this work, we tested a more novel, deep learning architecture designed to generate synthetic wearable and mood data using conditional GANs and multi-task learning. Our results suggest that these GANs can generate realistic wearable sensing and mood data, but do not improve downstream mood prediction models.

Table 2: Area under the receiver operating curve (AUROC) of the downstream mood classification models, split by the augmentation method (each row) and cross-validation split (each column).

Method	20%	40%	60%	80%
No Augmentation	0.5715	0.5822	0.5899	0.5934
Jittering	0.5727	0.5802	0.5861	0.5910
Scaling	0.5747	0.5825	0.5905	0.5952
Random Scaling	0.5739	0.5819	0.5885	0.5950
GAN	0.5032	0.5177	0.5391	0.5638
Multi-task CGAN	0.5233	0.5529	0.5734	0.5812

These findings suggest two important insights. First, the quality of synthetic data generation affects the quality of classification performance. The multi-task CGAN appeared to generate higher quality synthetic data than the simpler GAN model, and this higher quality synthetic data improved mood prediction, relative to the simpler GAN. But, performance was still reduced compared to the simpler methods, suggesting that scaling and jittering does not drastically change the training data distribution, and does not affect mood prediction.

Second, even with realistic samples, these initial findings suggest that these data augmentation methods do not improve mood prediction. Other recent work has shown a similar poor generalization accuracy of these mental health prediction models in large samples [21, 33]. We had hoped that our cross-validation approach, which aimed to personalize models by retraining them across time with more individual-level data, would improve performance. This was not the case. Thus, the low generalization accuracy in this work suggests further complexities in modeling mental health from sensed-behaviors, across both participants and time [4].

In the future, we need to investigate why the generalization accuracy of these models is poor. Visualizing multivariate synthetic and actual data may better compare the data distributions and help us understand why the augmentation methods did not mood prediction. Further evaluating these results across different datasets, clusters (eg, different K-means starting points) cross-validation splits, generative and prediction models, and conducting ablation studies to understand the impact of specific modeling choices (eg, adding the conditional network) would help uncover if poor performance is attributed to specific methods in this work, or broader complexities modeling wearable and mental health outcomes data. For example, performance might be poor because it is difficult to model the varying relationships between behavior and mental health across individuals over time [4]. Future work could analyze if augmenting data specifically in less represented subgroups improves generalization accuracy. We look forward to discussing these challenges in the workshop.

ACKNOWLEDGMENTS

D.A. is funded by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139899, and a Digital Life Initiative Doctoral Fellowship. Data collection for the Intern Health Study is funded by NIMH Grant No. R01MH101459. Computing resources were funded by a Microsoft Azure Cloud Computing Grant through the Cornell Center for Data Science for Enterprise and Society.

REFERENCES

[1] Daniel A Adler, Dror Ben-Zeev, Vincent W-S Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. 2020. Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks. *JMIR mHealth and uHealth* 8, 8 (Aug. 2020), e19962. https://doi.org/10.2196/19962

[2] Daniel A. Adler, Vincent W.-S. Tseng, Gengmo Qi, Joseph Scarpa, Srijan Sen, and Tanzeem Choudhury. 2021. Identifying Mobile Sensing Indicators of Stress-Resilience. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (June 2021), 51:1–51:32. https://doi.org/10.1145/3463528

[3] Daniel A. Adler, Fei Wang, David C. Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE* 17, 4 (April 2022), e0266516. https://doi.org/10.1371/journal.pone.0266516 Publisher: Public Library of Science.

[4] Daniel A. Adler, Fei Wang, David C. Mohr, Deborah Estrin, Cecilia Livesey, and Tanzeem Choudhury. 2022. A call for open data to develop mental health digital biomarkers. *BJPsych Open* 8, 2 (March 2022). https://doi.org/10.1192/bjo.2022.28 Publisher: Cambridge University Press.

[5] DeWitt C. Jr Baldwin, Steven R. Daugherty, Ray Tsai, and Michael J. Jr Scotti. 2003. A National Survey of Residents' Self-Reported Work Hours: Thinking Beyond Specialty. *Academic Medicine* 78, 11 (Nov. 2003), 1154–1163. https://journals.lww.com/academicmedicine/Fulltext/2003/11000/A_National_Survey_of_Residents__Self_Reported_Work.18.aspx

[6] Theodore P. Beauchaine and Julian F. Thayer. 2015. Heart rate variability as a transdiagnostic biomarker of psychopathology. *International Journal of Psychophysiology* 98, 2, Part 2 (Nov. 2015), 338–350. https://doi.org/10.1016/j.ijpsycho.2015.08.004

[7] Rich Caruana. 1997. Multitask Learning. *Machine Learning* 28, 1 (July 1997), 41–75. https://doi.org/10.1023/A:1007379606734

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. https://doi.org/10.1145/2939672.2939785 arXiv:1603.02754 [cs].

[9] Fitbit. 2020. Fitbit Development: Reference. https://dev.fitbit.com/build/reference/ Library Catalog: dev.fitbit.com.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. Google-Books-ID: omivDQAAQBAJ.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. (June 2014). https://arxiv.org/abs/1406.2661v1

[12] Constance Guille, Heather Speller, Rachel Laff, C. Neill Epperson, and Srijan Sen. 2010. Utilization and Barriers to Mental Health Services Among Depressed Medical Interns: A Prospective Multisite Study. *Journal of Graduate Medical Education* 2, 2 (June 2010), 210–214. https://doi.org/10.4300/JGME-D-09-00086.1

[13] Shota Haradal, Hideaki Hayashi, and Seiichi Uchida. 2018. Biosignal Data Augmentation Based on Generative Adversarial Networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 368–371. https://doi.org/10.1109/EMBC.2018.8512396 ISSN: 1558-4615.

[14] Brian Kenji Iwana and Seiichi Uchida. 2021. An Empirical Survey of Data Augmentation for Time Series Classification with Neural Networks. *PLOS ONE* 16, 7 (July 2021), e0254841. https://doi.org/10.1371/journal.pone.0254841 arXiv:2007.15951 [cs, stat].

[15] Nicholas C. Jacobson and Brandon Feng. 2022. Digital phenotyping of generalized anxiety disorder: using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life. *Translational Psychiatry* 12, 1 (Aug. 2022), 1–7. https://doi.org/10.1038/s41398-022-02038-1 Number: 1 Publisher: Nature Publishing Group.

[16] David A. Kalmbach, Yu Fang, J. Todd Arnedt, Amy L. Cochran, Patricia J. Deldin, Adam I. Kaplin, and Srijan Sen. 2018. Effects of Sleep, Physical Activity, and Shift Work on Daily Mood: a Prospective Mobile Monitoring Study of Medical Interns. *Journal of General Internal Medicine* 33, 6 (June 2018), 914–920. https://doi.org/10.1007/s11606-018-4373-2

[17] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* 114, 1 (April 2009), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

[18] Xi'ang Li, Jinqi Luo, and Rabih Younes. 2020. ActivityGAN: generative adversarial networks for data augmentation in sensor-based human activity recognition. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp-ISWC '20)*. Association for Computing Machinery, New York, NY, USA, 249–254. https://doi.org/10.1145/3410530.3414367

[19] Tony Liu, Jonah Meyerhoff, Johannes C. Eichstaedt, Chris J. Karr, Susan M. Kaiser, Konrad P. Kording, David C. Mohr, and Lyle H. Ungar. 2022. The relationship between text message sentiment and self-reported depression. *Journal of Affective Disorders* 302 (April 2022), 7–14. https://doi.org/10.1016/j.jad.2021.12.048

[20] Douglas A. Mata, Marco A. Ramos, Narinder Bansal, Rida Khan, Constance Guille, Emanuele Di Angelantonio, and Srijan Sen. 2015. Prevalence of Depression and Depressive Symptoms Among Resident Physicians: A Systematic Review and Meta-analysis. *JAMA* 314, 22 (Dec. 2015), 2373–2383. https://doi.org/10.1001/jama.2015.15845 Publisher: American Medical Association.

[21] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. 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* 6, 4 (Jan. 2023), 176:1–176:32. https://doi.org/10.1145/3569483

[22] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual review of clinical psychology* 13 (May 2017), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

[23] Sandrine R. Müller, Xi (Leslie) Chen, Heinrich Peters, Augustin Chaintreau, and Sandra C. Matz. 2021. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports* 11 (July 2021), 14007. https://doi.org/10.1038/s41598-021-93087-x

[24] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research* 17, 7 (July 2015). https://doi.org/10.2196/jmir.4273

[25] Srijan Sen. 2021. Intern Health Study. https://doi.org/10.3886/E129225V1

[26] Srijan Sen, Henry R. Kranzler, John H. Krystal, Heather Speller, Grace Chan, Joel Gelernter, and Constance Guille. 2010. A prospective cohort study investigating factors associated with depression during medical internship. *Archives of General Psychiatry* 67, 6 (June 2010), 557–565. https://doi.org/10.1001/archgenpsychiatry.2010.41

[27] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 1 (July 2019), 60. https://doi.org/10.1186/s40537-019-0197-0

[28] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* (2017), 1–1. https://doi.org/10.1109/TAFFC.2017.2784832

[29] Vincent W.-S. Tseng, Akane Sano, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Marta Hauser, John M. Kane, Emily A. Scherer, Rui Wang, Weichen Wang, Hongyi Wen, and Tanzeem Choudhury. 2020. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Scientific Reports* 10, 1 (Sept. 2020), 15100. https://doi.org/10.1038/s41598-020-71689-1 Number: 1 Publisher: Nature Publishing Group.

[30] Terry Taewoong Um, Franz Michael Josef Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data Augmentation of Wearable Sensor Data for Parkinson's Disease Monitoring using Convolutional Neural Networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Nov. 2017), 216–220. https://doi.org/10.1145/3136755.3136817 arXiv: 1706.00527.

[31] Rui Wang, Emily A. Scherer, Vincent W. S. Tseng, Dror Ben-Zeev, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, and Michael Merrill. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*. ACM Press, Heidelberg, Germany, 886–897. https://doi.org/10.1145/2971648.2971740

[32] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. 2021. Time Series Data Augmentation for Deep Learning: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 4653–4660. https://doi.org/10.24963/ijcai.2021/631 arXiv:2002.12478 [cs, eess, stat].

[33] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (Jan. 2023), 190:1–190:34. https://doi.org/10.1145/3569485

[34] Han Yu and Akane Sano. 2022. Semi-Supervised Learning and Data Augmentation in Wearable-based Momentary Stress Detection in the Wild. http://arxiv.org/abs/2202.12935 arXiv:2202.12935 [cs, eess].