WatchAnxiety: A Transfer Learning Approach for State Anxiety Prediction from Smartwatch Data

Md Sabbir Ahmed[1] , Noah French[2], Mark Rucker[1], Zhiyuan Wang[1], Taylor Myers-Brower[2], Kaitlyn Petz[2], Mehdi Boukhechba[3], Bethany A. Teachman[2], Laura E. Barnes[1]

[1]Dept. of Systems & Information Engineering, [2]Dept. of Psychology, University of Virginia, USA
[3]Johnson & Johnson Innovative Medicine, USA
{msabbir, njf5cu, mr2an, vmf9pr, cdf3zf, kdp8y, bteachman, lb3dp}@virginia.edu; mboukhec@its.jnj.com

***Abstract*—Social anxiety is a common mental health condition linked to significant challenges in academic, social, and occupational functioning. A core feature is elevated momentary (state) anxiety in social situations, yet little prior work has measured or predicted fluctuations in this anxiety throughout the day. Capturing these intra-day dynamics is critical for designing real-time, personalized interventions such as Just-In-Time Adaptive Interventions (JITAIs). To address this gap, we conducted a study with socially anxious college students (N=91; 72 after exclusions) using our custom smartwatch-based system over an average of 9.03 days (SD = 2.95). Participants received seven ecological momentary assessments (EMAs) per day to report state anxiety. We developed a base model on over 10,000 days of external heart rate data, transferred its representations to our dataset, and fine-tuned it to generate probabilistic predictions. These were combined with trait-level measures in a meta-learner. Our pipeline achieved 60.4% balanced accuracy in state anxiety detection in our dataset. To evaluate generalizability, we applied the ~~trained~~ training approach to a separate hold-out set from the TILES-18 dataset—the same dataset used for ~~TL~~ pretraining. On 10,095 once-daily EMAs, our method achieved 59.1% balanced accuracy, outperforming prior work by at least 7%.**

I. Introduction

Social anxiety, or anxiety tied to social situations in which one may be evaluated negatively, is a prevalent mental health problem. An estimated 12.1% of individuals in the U.S. meet the criteria for social anxiety disorder at some point in their life [1]. Social anxiety often limits individuals' lives and is associated with avoiding potentially meaningful careers that require social interactions, avoiding romantic relationships, and delaying starting families [2]. Existing research shows that helping people respond to state anxiety in more effective ways (e.g., by challenging anxious thinking and approaching rather than avoiding feared situations) can reduce overall levels of social anxiety [3]. However, much of the existing research using passive sensing to detect anxiety has focused on predicting between-person differences in anxiety levels—most commonly trait anxiety, a stable and enduring tendency to experience anxiety across time and situations (e.g., [4])—or general anxiety symptoms [5]. While predicting trait anxiety through passive sensing can be useful for early identification of mental health conditions, advancing toward the detection of within-person fluctuations in anxiety (i.e., state anxiety) is essential for enabling real-time, adaptive interventions that address anxiety in the moment.

Past studies have used passive sensing to predict within-person anxiety level at the daily timescale (e.g., [6]), but only a handful of studies (e.g., [7]) have attempted to estimate within-person fluctuations in anxiety measured at the timescale of hours or minutes, with most such research conducted in controlled laboratory settings. Only one study to our knowledge has attempted to predict within-person fluctuations in anxiety measured multiple times per day outside of a controlled lab setting [8]. However, in [8], the authors used R^2 as the evaluation metric, which does not directly reflect predictive accuracy, leaving the model's effectiveness in identifying moments of state anxiety unclear. Moreover, their models relied on smartphone sensor data, which may be less effective for detecting momentary anxiety, as smartphones are not always carried as frequently as smartwatches [9].

For our study, we developed WatchAnxiety, a smartwatch-based system that advances wearable computing and pervasive health by using transfer learning to predict state anxiety. Validated on 2,742 real-world EMA responses, the model achieved 60.4% balanced accuracy and F1 score. To assess generalizability, we then applied our meta-learning approach to an independent dataset of 10,095 state-anxiety EMAs—bringing the total labeled samples to over 12,000, the largest evaluation to date. This scale is noteworthy given that wearable mental-health research is often hampered by limited labeled data, which can impede robust validation and real-world deployment.

II. Methodology

A. Dataset Construction

1) System Design: We developed a smartwatch system (expected to be compatible with any Wear OS-based smartwatch) for real-time collection of physiological, behavioral, and acoustic data. To protect participants' privacy, data collection is disabled between 12 AM and 8 AM and automatically pauses when the watch is removed. The system operates on a 5-minute duty cycle, capturing data for 1 minute per cycle. Data are uploaded to secure Amazon S3 storage either manually (via button press) or automatically when the watch is charging and connected to Wi-Fi. Although the system supports multiple sensing modalities, this study focuses exclusively on heart rate (HR)—a widely available physiological marker on commercial smartwatches with strong relevance to health monitoring.

2) Participants: All study procedures were approved by IRB of the University of Virginia (UVA). We recruited par-

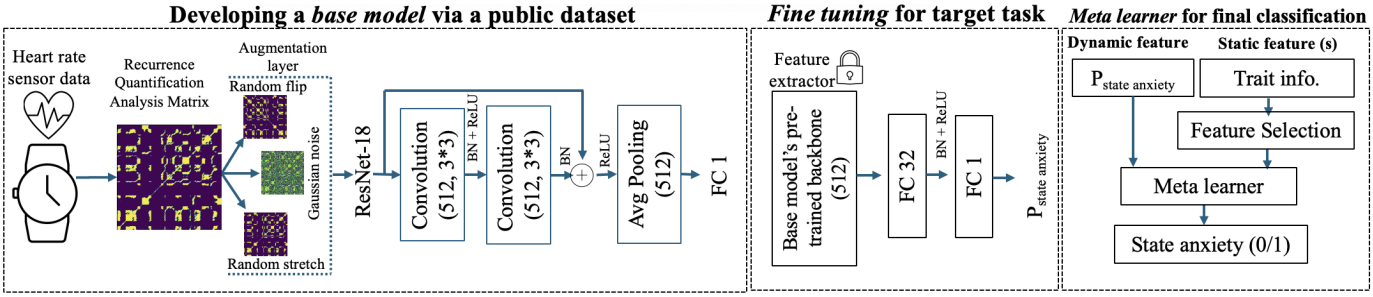

Fig. 1: WatchAnxiety system for identifying state anxiety. FC: Fully Connected, BN: Batch Normalization.

ticipants with moderate-to-severe levels of social anxiety, operationalized as a score of at least 34 on the Social Interaction Anxiety Scale (SIAS) following prior research [10]. Each participant was provided with either a Samsung Galaxy Watch 5 or 5 Pro pre-installed with our system. Moreover, participants installed the Sensus [11] app on their personal smartphone to receive the EMA surveys. It is worthwhile to mention that early participants received an earlier version of our on-watch system with a duty cycle around 10 minutes, while later participants mostly used the improved version with a 5-minute cycle.

We recruited 91 undergraduate students at UVA. Nineteen participants were excluded due to limited data from the initial version of the system, early withdrawal, or missing watch data within the analysis window. The final sample included 72 participants used for model development.

3) Baseline Data Collection: During the initial study visit, participants completed baseline surveys that captured trait-level mental health characteristics: SIAS, Brief Fear of Negative Evaluation (BFNE), Difficulties in Emotion Regulation (DERS), Depression, Anxiety, and Stress Scales-21 (DASS-21), Adult Rejection Sensitivity Questionnaire (A-RSQ), and Cambridge Depersonalization Scale 2-item version (CDS-2). For missing items (57 item-level responses; 0.59%), we imputed missing values using the mean of the remaining valid responses of the participant on the corresponding scale. For reverse-scored items, reverse coding was applied prior to both imputation and scale score aggregation. Importantly, imputation was performed independently for each participant, using only their own responses, thereby avoiding the use of data from other participants and preventing data leakage.

4) Measuring State Anxiety: Following the initial study visit, participants reported their state anxiety via ecological momentary assessments (EMAs) delivered through the Sensus smartphone app up to seven times daily for 10 days. EMAs were randomly scheduled every two hours between either 8 AM–10 PM or 10 AM–12 AM, based on participant preference. Each EMA asked, "I feel...," with responses recorded on a slider from 1 ("not at all anxious") to 10 ("very anxious"). For classification, responses were binarized: a rating of 1 was coded as class 0 (no anxiety) and all values greater than 1 as class 1 (any level of anxiety).

Among the 72 participants, HR data were available for 650 total participant-days (mean = 9.03 days, SD = 2.95), with a total of 3,663 state anxiety EMA responses (mean = 50.88 EMAs, SD = 15.91). However, the number of EMA responses included for model development varied based on the explored time window. Specifically, at least 50 HR samples were available within the 1-hour, 1.5-hour, and 2-hour windows for 74.94%, 75.87%, and 76.41% of the EMA responses, respectively. Selecting the appropriate window therefore is critical: larger windows improve data availability but increase overlap between EMAs, while shorter windows may better capture transient physiological markers relevant to state anxiety but increase the chance for data missingness within windows. To balance these trade-offs, we set a 50-sample threshold—an empirically supported cutoff, as nearly all sensor probe start times produced at least 50 HR readings within a one-minute period at ~1 Hz sampling. In other words, if at least one probe occurred within the relevant time window prior to an EMA submission, we included that EMA for model development.

B. Model Development

1) Feature Space: To construct the input feature space, we first estimate R-R intervals (RRI) from HR using the formula $\text{RRI} = \frac{60}{\text{HR}}$ [12]. We also excluded HR values outside the physiologically plausible range: above the age-adjusted maximum (220 ~~minus~~ - age) and below 40 bpm, a threshold reflecting the resting HR of very fit individuals. We then estimated the corresponding RRI timestamps using the cumulative sum of the RRI values, consistent with implementations in widely used packages (e.g., NeuroKit2). Using the inferred RRI and the corresponding timestamps, we performed a recurrence quantification analysis (RQA) of HR variability using the NeuroKit2 package **Makowski˙NeuroKit2˙2021**. To take advantage of pre-trained ResNet-18 in our base model, we adopted an image-based approach by transforming HR into recurrence plots. These plots ~~encode~~ are based on time-delayed embeddings of physiological signals, revealing dynamic patterns that can be useful for predicting anxiety.

2) Transfer Learning Approach: Transfer learning (TL) has shown promise across diverse prediction tasks and is particularly beneficial in scenarios with limited data [13]. Given our relatively small sample size (N = 72), TL is well-suited. To develop the base model for TL, we used the TILES-18 dataset [14], which includes sensor data from 212 hospital workers collected via multiple devices, including Fitbit. Each day, participants responded to a state anxiety EMA "Please select the response that shows how anxious you feel at the moment"

on a scale of 1 to 5. We used the same approach (section II-A4) as used for our dataset to create the target variable for classification task. After pre-processing and filtering for entries with at least 50 HR samples, a total of 10,278 EMA responses were available for modeling, with 38.35% labeled as class 1 and 61.65% as class 0. However, to explore generalization (section III-A), the dataset was reduced to 10,095 EMAs due to missing aggregated trait scores in TILES-18 for some participants.

For the model, we adopted ResNet-18 [15] without its classification head and initialized it with ImageNet-pretrained weights to leverage transferable representations. We added a residual block, a global average pooling layer, and a final output layer with a single neuron to predict the probability of state anxiety. To address class imbalance, we applied class weights and used the weighted binary focal loss as the loss function. The model was trained with an SGD optimizer with a learning rate of $1e-4$ and trained for a maximum of 20 epochs to obtain reasonable weights for the new layers. We then restored the weights corresponding to the lowest validation loss and fine-tuned the entire model using a reduced learning rate of $1e-8$. To prevent overfitting, early stopping was applied if validation loss did not improve for 3 consecutive epochs.

To reduce computational overhead, we employed a leave-five-out cross-validation (LFOCV) strategy, holding all state anxiety responses of five participants for testing in each fold. Of the remaining participants, two were selected for validation (i.e., used for model selection and early stopping) while the rest were used for training. In some cases, a validation participant reported only one class of state anxiety, which could bias the model. To address this and improve generalization, validation participants were chosen, when available, such that the ratio of class 1 to class 0 was within 10% of that in the training set.

3) Model Tuning: A key difference between the TILES-18 dataset, used for base model development, and our target task lies in the EMA protocol: TILES-18 collected state anxiety once a day, while our study administered EMA seven times a day to capture intraday fluctuations. TILES-18 also relied on Fitbit devices with continuous data collection, while our custom system employed a duty-cycled sampling strategy to support real-time processing and conserve battery life - enabling both data collection and future on-device interventions. Furthermore, the study populations differed: TILES-18 involved hospital shift workers, while our participants were undergraduate students.

Since the base models were trained using LFOCV on TILES-18 data, multiple models were generated. We selected the one with the highest balanced accuracy in its respective test set. To adapt the model, we removed the top classification layer and used the 512-neuron global average pooling layer as output. We then added a 32-unit fully connected layer, followed by batch normalization and a ReLU activation function. A final dense layer with a single neuron produced the predicted probabilities. This architecture was inspired by the squeeze-and-excitation (SE) block from SENet, where a low-dimensional representation is learned post-global pooling to enhance generalizable feature learning. The weights of the base model were frozen and the new layers were trained using the Nadam optimizer (learning rate = $1e-5$), as SGD yielded suboptimal results in this context. To avoid overfitting or underfitting, we employed a custom callback that restored the best model based on training dynamics. Specifically, we allowed up to a 3% tolerance between training and validation loss, restoring the weights with the smallest difference if the validation loss dropped below the training loss or exceeded the tolerance. Training, validation, and testing followed the same LFOCV protocol described in Section II-B2.

4) Meta-learner Development and Evaluation: After generating predicted probabilities from the fine-tuned model, we incorporated trait measures (Section II-A3) to train a meta-learner. This approach is practical for real-world deployment, as trait assessments need to be completed only once prior to system use. To retain only relevant characteristics, we applied feature selection based on information gain to identify the most predictive traits. We then trained lightweight classifiers, K-Nearest Neighbors (KNN), Logistic Regression (Logit), and Decision Tree, as meta-learners for classifying state anxiety. The use of lightweight models was motivated by the goal of minimizing overfitting.

Given that our meta-learner is lightweight and fast to train, we applied leave-one-out cross-validation (LOOCV) at the final evaluation stage. This setup, in which the intermediate stage uses LFOCV and the final stage uses LOOCV, avoids information leakage. In contrast, using LFOCV in the final stage and LOOCV earlier could introduce leakage by allowing a participant seen during training in intermediate stage to possibly reappear in the test sets of the meta-learner. For model evaluation, we report balanced accuracy, precision, recall, F1-score, and specificity. To address class imbalance, both precision and F1-score were computed as weighted metrics.

III. Results and Discussion

Although we explored 3 meta-learners, Logit consistently performed better; thus, we report results for Logit only. Across the windows evaluated, the performance was relatively similar (Table I); however, the 1.5~~hour~~ hour window offered a favorable balance between recall (~~58. 1~~58.1%) and balanced accuracy (~~60. 4~~60.4%). To assess robustness, we compared it with two baseline models. Baseline 1 is a model based on all trait measures, while baseline 2 is a random classifier with uniform probability across both classes. As shown in Table I, our ~~meta- learner~~ meta-learner using a 1.5-hour window outperformed both baselines. Though baseline 1 has a comparable balanced accuracy (60.4% vs. 57.3%) with our meta learner, empirically, we found a model based on trait measures predicted always either class 1 or 0 for all days of each participant (section III-B for details).

TABLE I: Performance of baseline models and our meta-learner. BA = Balanced Accuracy, Prec = Precision, Rec = Recall, Spec = Specificity.

Model	EMAs	Trait	Prec.	Rec.	Spec.	BA	F1
Meta (1h)	2703	4	61.8	57.9	61.9	59.9	60.0
Meta (1.5h)	2742	4	**62.3**	**58.1**	62.7	**60.4**	**60.4**
Meta (2h)	2765	4	62.6	56.6	**64.8**	60.7	60.3
Baseline 1	2765	5	59.2	59.6	55.1	57.3	58.3
Baseline 2	2765	–	51.2	44.4	53.3	48.9	48.4

A. Approach Generalization

To assess the generalizability of our modeling pipeline and benchmark it against existing methods, we conducted external validation using the TILES-18 dataset [14]. We compared our approach to a prior study [16] that predicted state anxiety using features derived from Fitbit and other devices. Since our model relies on watch-sensed heart rate (HR), we implemented two baseline versions: one using all Fitbit-derived features (e.g., cardio, fat burn, steps, sleep) and another using only HR features, as in the original study. As shown in Table II, our meta-learner model substantially outperformed both baselines—for instance, achieving a 7.9% higher balanced accuracy than the all-feature model.

TABLE II: TILES-18 ~~Evaluation~~evaluation. ~~BA~~ F = ~~Balanced Accuracy, Spec = Specificity~~Features.

Model	EMAs	F.	Prec.	Rec.	Spec.	BA	F1
Meta-learner (Ours)	10095	7	**61.4**	**50.9**	**67.4**	**59.1**	**61.3**
HR-only [16]	10095	1	51.3	3.6	95.7	49.7	49.1
All Fitbit [16]	10095	25	54.4	23.7	78.8	51.2	54.7

B. Ablation Study

To evaluate the contributions of the transfer learning (TL) model and the meta-learner, we conducted an ablation study. First, we assessed the TL model alone—without the meta-learner—on both our dataset and the external TILES-18 dataset. On our dataset, the TL model achieved 58.5% recall and 36.7% specificity. On TILES-18, it achieved 38.5% recall and 58.9% specificity. Despite lower overall performance, the TL model outperformed previously published approaches [16] in recall when evaluated on 10,095 EMA responses from TILES-18. As shown in Table II, recall for the prior models ranged from just 3.6% (HR-only features) to 23.7% (all Fitbit features), compared to 38.5% with our TL-only model and 50.9% with meta-learner.

We trained a trait-only model using four selected trait measures, applying the same feature selection, classifier (Logit), and hyperparameters as the meta-learner. While overall metrics were comparable (e.g., BA: 60.4% vs. 60.16%; F1: 60.4% vs. 59.2%), the trait-only model failed to capture intra- and inter-day variability, consistently predicting the same class per participant (BA: 0%, 50%, or 100%). In contrast, our meta-learner produced more temporally sensitive, participant-specific predictions. For example, in our dataset, 7 participants had per-participant BA between 50%–100% and 9 between 0%–50%, with similar results on the TILES-18 dataset. These differences reflect the static nature of trait-only inputs versus the temporal variation in TL-derived probabilities used by the meta-learner.

IV. Conclusion and Future Work

We propose a model that leverages ~~smartwatch-sensed~~ watch-sensed data to predict state anxiety. While it outperforms baseline models, there is still considerable room for improvement to support more precise and personalized interventions. Future work will explore incorporating additional sensor modalities and enabling on-device detection of timely intervention opportunities. In parallel, strategies such as thresholds jointly determined by clinicians and users (e.g., lowering the threshold to increase sensitivity) could help balance sensitivity and specificity, thereby mitigating potential negative effects. Also, our deep learning–based approach is limited by interpretability, which future work could address.

References

[1] R. C. Kessler et al., "Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States," *International journal of MPR*, 2012.

[2] A. Caspi, G. H. Elder Jr., and D. J. Bem, "Moving away from the world: Life-course patterns of shy children," *Developmental Psychology*, 1988.

[3] R. Kindred, G. W. Bates, and N. L. McBride, "Long-term outcomes of cognitive behavioural therapy for social anxiety disorder: A meta-analysis of randomised controlled trials," *Journal of Anxiety Disorders*, 2022.

[4] M. Boukhechba et al., "Predicting Social Anxiety From Global Positioning System Traces of College Students: Feasibility Study," *JMIR Mental Health*, Jul. 2018.

[5] N. K. Sahu, S. Gupta, and H. Lone, "Wearable Technology Insights: Unveiling Physiological Responses During Three Different Socially Anxious Activities," *ACM JCSS*, Jun. 2024.

[6] H. Rashid et al., "Predicting Subjective Measures of Social Anxiety from Sparsely Collected Mobile Sensor Data," *ACM IMWUT*, 2020.

[7] M. A. Larrazabal et al., "Understanding state social anxiety in virtual social interactions using multimodal wearable sensing indicators," *IEEE SMARTCOMP'25 [In Press]*, 2025.

[8] N. C. Jacobson and S. Bhattacharya, "Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments," *Behaviour Research and Therapy*, Feb. 2022.

[9] F. Shahmohammadi et al., "Smartwatch based activity recognition using active learning," in *2017 IEEE/ACM CHASE*, IEEE, 2017.

[10] E. R. Toner et al., "Wearable Sensor-based Multimodal Physiological Responses of Socially Anxious Individuals in Social Contexts on Zoom," *IEEE Transactions on Affective Computing*, 2025.

[11] H. Xiong et al., "Sensus: A cross-platform, general-purpose system for mobile crowdsensing in human-subject studies," *Proc. of the UbiComp'16*, 2016.

[12] E. S. Prakash and Madanmohan, "How to tell heart rate from an ecg? (learning objects 769 and 878)," *Advances in Physiology Education*, 2005.

[13] A. Ebbehoj et al., "Transfer learning for non-image data in clinical research: A scoping review," *PLOS Digital Health*, Feb. 2022.

[14] K. Mundnich et al., "Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, 2020.

[15] K. He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, Jun. 2016.

[16] R. Pranjal et al., "Toward privacy-enhancing ambulatory-based well-being monitoring: Investigating user re-identification risk in multimodal data," *ICASSP'23*, 2023.

Dear program chairs,

We thank the reviewers for their time and insightful comments. We also thank you for giving us an opportunity to address the raised issues. We addressed those in the revised manuscript and uploaded it as the camera-ready version. Here, we are responding to each of the concerns raised by the reviewers.

1. Response to Reviews of Reviewer B6YB

1) Given 60% accuracy in a binary classification task, which is 10% above random guessing, how practical is this technology when sensing and designing interventions based on it?

Our response: We appreciate this important question. An around 10% improvement over random guessing is meaningful given the scale of our evaluation—over 12,000 EMA samples—representing the largest evaluation of state anxiety prediction to date (previous studies [1], [2] tested on $<$1000 EMA samples). Nonetheless, we acknowledge that the performance may still limit real-world impact, which we have noted in the "Conclusion and Future Work" section: "While it outperforms baseline models, there is still considerable room for improvement to support more precise and personalized interventions. Future work will explore incorporating additional sensor modalities and enabling on-device detection of timely intervention opportunities." In section IV of the revised manuscript, we note that strategies such as thresholds jointly determined by clinicians and users (e.g., lowering the threshold to increase sensitivity) could help balance sensitivity and specificity and mitigate potential negative effects.

2) There should be some rationale on why other physiological signals (like Electrodermal activity or EDA, respiratory rate) have not been used in addition to HR. Adding them might have improved the classification results further.

Our response: We appreciate the reviewer's suggestion to consider additional physiological signals such as EDA and respiratory rate. Our initial development and rigorous testing were conducted on Galaxy Watches, which, to the best of our knowledge, do not provide EDA or respiratory rate sensing. We intentionally focused on heart rate to evaluate system performance since it is the most common physiological sensor available on commercial smartwatches, as stated in our manuscript: "Although the system supports multiple sensing modalities, this study focuses exclusively on heart rate (HR)—a widely available physiological marker on commercial smartwatches with strong relevance to health monitoring."

3) Is the Samsung 5/5 pro at the core of the developed smartwatch system? If so, it must be clearly stated. Or, if the developed sensor system is independent of the Samsung watch, why it was given to the participants?

Our response: Thank you for raising this point. The Samsung Galaxy Watch 5/5 Pro is not at the core of the developed system. Our system is platform-independent and expected to be compatible with any Wear OS–based smartwatch. We added language to the "System Design" section of the revised manuscript to clarify this. While the Galaxy Watch 5/5 Pro was used for participant deployment, we also tested the system on the Galaxy Watch 6 and conducted initial testing on the Google Fossil Watch. We selected the Samsung watch for participants because our system had been rigorously validated across multiple Galaxy models and also offered a significantly larger battery capacity (e.g., Watch 5: $\sim$410 mAh, Watch 5 Pro: $\sim$590 mAh vs. Google Pixel Watch 2: $\sim$306 mAh, the latest model available during the study), which was essential for enabling full-day continuous sensing with a single daily charge.

4) Since the TILES-18 data was collected using Fitbit, which has a different sensing hardware and calibration system, just the transfer learning followed by the meta learning is not enough to ensure cross domain generalization/adaptation and resulted in underperforming accuracy. My assumption is that the transfer learning + meta learning approach is good for cross-population adaptation (fine tuning a model learned for hospital workers with student data) but not for cross deice adaptation.

Our response: We thank the reviewer for this insightful comment. We did not directly use the base model trained on Fitbit-derived heart rate data to predict state anxiety in our dataset. Instead, we fine-tuned the base model with a updated architecture (Figure 1) using our own dataset, which may help with cross-device adaptation—similar to how transfer learning followed by meta learning may support cross-population adaptation— especially, given

that both Fitbit and Galaxy watches measure heart rate. That said, we agree that further investigation is needed to rigorously evaluate the extent to which this approach enables cross-device generalization, and we acknowledge this as an important direction for future work. However, we could not discuss this in the revised paper due to page limitations.

5) Earlier participants received a smartwatch system with 10-minute duty-cycle and the later ones received devices with 5-minute duty-cycle. What processing was done to ensure data coherence?

Our response: Our primary focus was to train a deep learning model capable of learning robust feature representations regardless of the number of available samples. The recurrence plots used in our approach are not affected by the sampling interval because we computed RRI timestamps using the cumulative sum of the RRI values, which depends on the RR interval values rather than the time spacing of consecutive heart rate samples, as stated in the manuscript: "We then estimated the corresponding RRI timestamps using the cumulative sum of the RRI values, consistent with implementations in widely used packages (e.g., NeuroKit2)". Consequently, the 5-minute or 10-minute duty cycle should not affect the derived RRI time. Moreover, the base model was trained on the TILES-18 dataset, which itself contains heart rate data sampled at non-uniform intervals ranging from approximately 5 seconds to 15 minutes, depending on participants' physical activity levels [3]. Developing a base model on such non-uniform data likely helped the model's ability to generalize to variable sampling intervals during fine-tuning and evaluation.

6) No details were given on how the HR signals are transformed into recurrence plots.

Our response: Thank you for pointing this out. In the revised manuscript (section "Model Development"), we now specify that the recurrence plots were generated using the neurokit2 package, a widely used toolkit for physiological signal processing.

2. Response to Reviews of Reviewer 5Tte

1) Although the paper converts anxiety levels into a binary classification task, the model's overall performance remains relatively modest, which may limit its practical utility.

Our response: Kindly see the response to 1st concern of reviewer B6YB.

2) The first baseline (trait-only) model is not sufficiently described in terms of implementation and purpose, and the second baseline (random classifier) does not offer meaningful comparative value. The paper lacks a rigorous comparison with existing deep learning–based or classical approaches for state anxiety detection, which makes it difficult to assess the model's relative performance in the broader literature.

Our response: We appreciate this valuable feedback. In addition to the two baseline models described, we also compared our approach with the pipeline from a prior study [2], providing a point of reference against existing work. Broader comparisons with both deep learning–based and classical approaches may offer a more rigorous assessment of our model's relative performance, and we hope to explore this in the future.

3) Despite external testing, both datasets are drawn from specific populations (college students and hospital workers), and it remains unclear whether the system generalizes to more diverse cohorts (e.g., older adults, adolescents, clinical populations).

Our response: We appreciate the reviewer's question on generalizability, as this is indeed a critical challenge for achieving real-world impact. As noted in the final paragraph of the Introduction section, our evaluation on more than 12,000 EMA samples represents, to our knowledge, the largest study of state anxiety prediction to date. This scale is noteworthy given that wearable mental-health research is often hampered by limited labeled data, which

can impede robust validation and real-world deployment. In comparison, prior studies [1], [2]—including the most recent work [1]—evaluated their models on fewer than 1,000 EMA samples. We see our work as an important step toward understanding generalization and view evaluation on more diverse cohorts (e.g., older adults, adolescents, clinical populations) as one of the directions for future research. As a first step, however, we believe healthcare workers and undergraduate college students are worthwhile populations for which to design anxiety-sensing systems. Healthcare workers have high rates of mental health problems, including burnout and anxiety, with some evidence suggesting that there have been long-term negative mental health impacts of the COVID-19 pandemic on this population [4]. Additionally, college-age individuals have high rates of anxiety relative to other age groups [5].

4) While the system design is technically sound, the paper does not discuss computational requirements, real-time feasibility, or energy efficiency—critical issues for practical deployment on wearable devices.

Our response: Thanks for pointing out practical deployment. Due to the page limitations of only 4 pages, we could not explore and include the computational requirements.

3. Response to Reviews of Reviewer dC3s

1) Limited sensor modalities: The study uses only heart rate despite having access to additional sensing modalities. Incorporating others (e.g., accelerometer, skin temperature) could boost predictive performance and robustness.

Our response: We appreciate this observation and agree that incorporating additional sensing modalities can further improve predictive performance and robustness. We have explicitly acknowledged this as a limitation in the "Conclusion and Future Work" section, where we state: "Future work will explore incorporating additional sensor modalities and enabling on-device detection of timely intervention opportunities."

2) Binary classification oversimplification: Binarizing anxiety into "none" vs. "any" potentially reduces the richness of data, especially given the original 1–10 scale. A regression or multi-class classification might preserve more information.

Our response: We appreciate this thoughtful comment. We agree that regression or multi-class classification could preserve more information and provide finer-grained predictions, which future work can explore. However, our binary prediction task ("none" vs. "any" anxiety) is still valuable for practical intervention scenarios, as it identifies moments when a user might need support, enabling timely and actionable just-in-time interventions. Indeed, just-in-time interventions largely rely on binary logic (e.g., don't deliver an intervention [0] vs. do deliver an intervention [1]), so our binarizing anxiety fits with what we see as the clearest application of our system. Furthermore, the one (to our knowledge) prior attempt at predicting state anxiety with passive sensing in a regression framework led to low performance [6] (i.e., models predicted 39% of variance in within-person anxiety fluctuations), indicating that substantial progress is still needed in this area.

3) Generalization scope: While performance on TILES-18 is promising, the datasets are demographically and behaviorally different. More diverse population studies (e.g., clinical samples, non-student populations) would improve ecological validity.

Our response: Kindly, see our response to the 3rd concern of the reviewer 5Tte.

4) Model interpretability: Although the architecture is well described, deeper insight into feature importance (e.g., SHAP or LIME analysis on the meta-learner) could improve interpretability, especially for healthcare deployment.

Our response: We appreciate this valuable suggestion. Our primary focus in this work was to develop a deep learning–based approach to achieve higher predictive performance compared to existing methods, which

comes with the trade-off of reduced interpretability. In section IV of the revised manuscript, we have explicitly acknowledged this limitation: ”our deep learning–based approach is limited by interpretability, which future work could address.”.

5) Privacy and deployment considerations: Discussion of privacy risks and the feasibility of on-device inference is minimal, especially relevant given continuous physiological monitoring.

Our response: We appreciate the reviewer's concern regarding privacy and deployment. We took several steps to respect participants' privacy. As stated in subsection ”System Design”: ”To protect participants' privacy, data collection is disabled between 12 AM and 8 AM and automatically pauses when the watch is removed”. Furthermore, our system does not continuously stream data; instead, it collects data for only one minute at predefined intervals (e.g., every five minutes), which further minimizes potential privacy risks.

In two different sections, we talked a bit about on-device deployment. In subsection ”Model Tuning”, we stated: ”TILES-18 also relied on Fitbit devices with continuous data collection, while our custom system employed a duty-cycled sampling strategy to support real-time processing and conserve battery life - enabling both data collection and future on-device interventions”. Also, in section IV, we stated: ”Future work will explore incorporating additional sensor modalities and enabling on-device detection of timely intervention opportunities.”

References

[1] M. A. Larrazabal, Z. Wang, M. Rucker, E. R. Toner, M. Boukhechba, B. A. Teachman, and L. E. Barnes, “Understanding state social anxiety in virtual social interactions using multimodal wearable sensing indicators,” *2025 IEEE International Conference on Smart Computing (SMARTCOMP)*, p. 162–169, 2025.

[2] R. Pranjal, R. Seshadri, R. Kumar Sanath Kumar Kadaba, T. Feng, S. S. Narayanan, and T. Chaspari, “Toward privacy-enhancing ambulatory-based well-being monitoring: Investigating user re-identification risk in multimodal data,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5, 2023.

[3] K. Mundnich, B. M. Booth, M. L'Hommedieu, T. Feng, B. Girault, J. L'Hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte, T. H. Falk, K. Lerman, E. Ferrara, and S. Narayanan, “Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers,” *Scientific Data*, vol. 7, Oct. 2020.

[4] J. Huang, Z.-T. Huang, X.-C. Sun, T.-T. Chen, and X.-T. Wu, “Mental health status and related factors influencing healthcare workers during the covid-19 pandemic: A systematic review and meta-analysis,” *PLoS One*, vol. 19, no. 1, p. e0289454, 2024.

[5] P. Jefferies and M. Ungar, “Social anxiety in young people: A prevalence study in seven countries,” *PloS one*, vol. 15, no. 9, p. e0239133, 2020.

[6] N. C. Jacobson and S. Bhattacharya, “Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments,” *Behaviour Research and Therapy*, vol. 149, p. 104013, 2022.