Don't Sleep on Sleep Data: Influence of Sleep Physiological Signals on Stress Detection

Soundarya Ramesh[1] **Takahiro Masuda**[2] **Hyung Woon Lee**[1]
Yongquan Hu[1] **Suranga Nanayakkara**[1]
[1]Augmented Human Lab, National University of Singapore
[2]Asahi Quality & Innovations, Ltd.
soundarya@ahlab.org, takahiro.masuda@asahi-qi.co.jp,
{leehw, yongquan, suranga}@ahlab.org

Abstract

Stress is a critical determinant of both short-term well-being and long-term health. While wearable sensors have enabled continuous monitoring of stress through physiological signals, existing approaches that rely only on *current physiology* have shown limited success. Prior work suggests that the *previous night's sleep* is predictive of stress, yet current methods typically use only *coarse sleep summaries* (e.g., duration, resting heart rate). In this paper, we argue that *fine-grained sleep physiological data* can provide richer insights for stress detection. We collect a month-long smartwatch dataset comprising both day-time and night-time physiological signals, including detailed sleep-derived features, and train two models – XGBoost and a custom multi-modal neural network. Our results provide initial evidence that incorporating fine-grained sleep features significantly improves stress detection, opening up several promising directions for future research.

1 Introduction

Stress is an important determining factor of a person's overall health and well-being. Reliable stress detection enables timely interventions to improve one's focus and decision-making abilities, while also contributing to early detection of long-term health disorders [12, 17]. Self-reported questionnaires are considered the gold standard for assessing perceived stress [32], but they are burdensome for users and impractical for continuous monitoring. An alternative is to leverage *physiological markers* such as heart rate variability (HRV), skin conductance, skin temperature, and respiration [14, 1, 31]. With the growing ubiquity and improved sensing accuracy of wearable devices[13, 8], many recent studies have explored continuous stress detection in real-world settings using wearable signals [38, 20, 9, 15]. However, these approaches have only achieved limited success in the wild[1], hinting at the need for additional sources of predictive information.

One promising but underexplored direction is the role of *sleep*. Prior work has shown that the previous night's sleep strongly predicts current-day stress and affect [27, 21]. Brink *et al.* [34] demonstrate that subjective sleep metrics, such as self-reported sleep quality and disturbances, significantly influence next-day affect. Joubert *et al.* [19] further found that objective sleep-derived measures, e.g., heart rate and high-frequency HRV from ECG, affect stress reactivity in cognitive tasks the following day.

Despite such evidence, current wearable-based stress detection methods typically reduce sleep information to coarse summaries (e.g., sleep duration, resting heart rate) [33, 18, 2, 30]. In this paper, we argue that *fine-grained sleep physiological features* from the previous night can provide richer insights into stress. We investigate this hypothesis using a newly collected dataset that combines

[1]Please refer to Appendix A for a detailed literature review.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

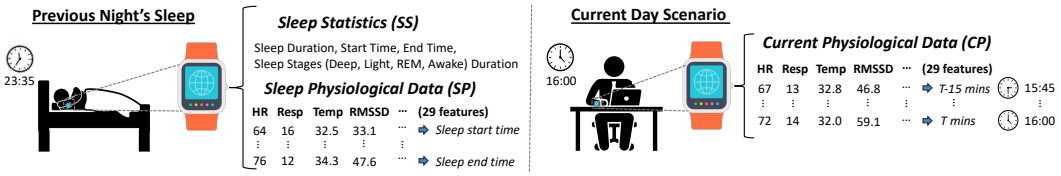

Figure 1: Figure illustrates the physiological features and sleep statistics extracted from sleep, as well as physiological features from user's current context during the day.

daytime physiology with detailed sleep-derived features, and evaluate the benefits of incorporating sleep information for improving stress detection.

Our Contribution. Unlike existing methods, we extract *fine-grained sleep features* to be used together with current physiological data for stress detection. Specifically, during sleep, we extract a total of 29 features, which includes time-domain, frequency-domain and non-linear HRV features from the PPG, skin temperature and accelerometer sensors, in addition to seven sleep statistics. Given the lack of public datasets that contain such data during day and night (i.e., including sleep), we collect our own dataset using a smartwatch, along with several self-reported labels per-day. Subsequently, we train two models – XGBoost and a custom multi-modal network – to demonstrate the effectiveness of sleep physiological data towards stress detection. Overall, our results provide initial evidence of sleep features significantly improving stress detection performance.

2 Method

2.1 Dataset

We collect physiological data through the Garmin Venu 3S smartwatch from 44 participants over 28 days (24 hours/day), as well as self-reported stress labels four times a day. Below, we provide details on the collected sensor data as well as the self-reported labels.

Physiological Signals. Through the Fitrockr API [10], we obtain several derived features from three sensors – photoplethysmography (PPG), skin temperature and accelerometer sensors – including respiration rate, heart rate, Spo2, beat-to-beat interval (BBI), skin temperature, as well as step count. All the data, except BBI, has one reading per minute. The BBI data consists of a continuous stream of time intervals between successive heart beats from which we extract heart rate variability (HRV) features. We compute HRV features in a moving window of five minutes, with a stride of a minute, to compute time-domain, frequency-domain as well as non-linear features using NeuroKit2 [23, 4]. This results in a total of 82 features, which we reduce to 29 features based on Spearman correlation (filtering those with correlation value beyond ± 0.8), based on the entire dataset. Please refer to Appendix Table 3 for the 29 selected features.

Sleep Statistics. We utilize sleep summary information such as total sleep duration, start and end times, as well as fraction of time spent in the sleep stages, i.e., light, deep, REM and awake stages.

Self-Reported Stress Labels. Through Experience Sampling Method (ESM), participants are prompted throughout the day and asked to log their stress once in each of the four daily sessions, namely morning, afternoon, evening and night[2]. Participants report stress value as an integer between 0 (not at all) and 100 (very much)[29]. For our analysis, we *binarize* this data, where values, < 50 and ≥ 50, are considered to represent *low* stress and *high* stress, respectively.

We perform supervised learning, hence we only utilize sensor data in the current context (i.e., 15 minutes of data until the reported label's timestamp) as well as the sleep data. To handle the missing sensor data samples that can occur due to improper watch fit or motion artifacts, we perform linear interpolation (in the forward direction for up to five minutes) per feature. Overall, after filtering samples with missing readings, we have a total of 2502 data samples, with a balanced label distribution of 1246 *low* stress and 1256 *high* stress samples (see Appendix Figure 2 for the label distribution).

[2]We segment the four sessions – morning, noon, evening and night, according to the following time intervals – 4:00-11:00, 11:00-15:00, 15:00-19:00, and 19:00-4:00, respectively.

2.2 Dataset Preparation

We utilize current physiological data (CP), sleep physiological data (SP) and sleep statistics (SS) for stress detection (Figure 1). For CP data, we consider a 15 minute context, similar to prior work [38], while for SP data, we consider the entire sleep duration, based on the sleep start and end times logged by the Garmin watch. For SS data, we utilize the seven sleep summary measures reported earlier. Below, we elaborate on feature extraction for the two machine learning models.

XGBoost. For the CP and SP data, we compute five statistics (mean, median, min, max, standard deviation) on each of the 29 features. By doing so on both CP and SP data, we obtain a total of 290 ($= 29 \times 5 \times 2$) features. Together with the SS data, this results in a total of 297 features per sample.

Custom Multi-Modal Network. In this case, our objective is to learn the temporal representation of the data through a multi-modal deep learning network. We keep the CP data as a two-dimensional matrix input (15 minutes $\times$ 29 features); however, for the SP data, we will segment the entire sleep duration due to the changing dimensions of the SP data caused by the inconsistent sleep duration. Specifically, we create 30 fragments such that each fragment is about 15 minutes (based on prior work [11]), and compute the mean statistic of features within each fragment, hence resulting in a two-dimensional matrix of fixed dimension (30 sleep fragments x 29 features) as input. The SS data was kept to a one-dimensional vector (7 features). All features were normalized with z-score normalization based on the mean and standard deviation values of the training data.

Cross Validation. We utilize time-series based cross validation method where each fold consists of 7 days, 2 days and 1 day of training, validation and test data, respectively. During every fold, we test on data from day N ($11 \leq N \leq 28$), while validating on data from two prior days, and training on the seven days prior to that. In this way, our training process is *causal* with the training and validation data always *preceding* the test data, thereby preventing data leakage.

2.3 Models

Based on prior work [38], we train an eXtreme Gradient Boosting (or XGBoost) model [3] using the statistical features mentioned in Section 2.2. We enumerate the hyperparameters used in Table 4.

We also train a custom multi-modal network that consists of two convolutional neural networks (CNN) encoders and one multilayer perceptron (MLP) encoder, for learning representations of the CP, SP and SS data, respectively (see Appendix Figure 3). Subsequently, we concatenate the representations of the three models and pass it through a final 1-layer MLP layer (with 32 neurons) for binary stress classification. For the CNN encoders, we utilize a 2-layer architecture with one-dimensional kernels, with the output channels of 16 and 32 respectively. For the MLP encoder, we similarly utilize a 2-layer architecture with the same output channels of 16 and 32. For all encoders, we include batch normalization, ReLU activation, and a dropout layer for regularization.

3 Experiments and Results

We evaluate the effect of incorporating previous night's sleep physiological features in overall stress detection (Section 3.1), as well as their effectiveness on detecting stress at different times of the day, i.e., morning, noon, evening and night (Section 3.2). We compute Area Under Receiver Operator Characteristics (AUROC) metric scores on six different feature combinations, involving presence or absence of current physiological (CP), sleep statistics (SS), and sleep physiological (SP) data.

3.1 Effect of Sleep on Overall Stress Detection

Tables 1 and 2 illustrate the overall performance (see column labeled 'All') of XGBoost and our multimodal network respectively, across the six feature combinations. All of the feature combinations with sleep physiological data (i.e., SP, SS+SP and CP+SS+SP) achieve a mean AUROC over 67%. Furthermore, they also achieve statistical significance[3], over the other feature combinations (i.e, CP, SS and CP+SS), demonstrating the importance of *fine-grained* sleep features for stress detection. In addition, Appendix Figure 4 depicts the top 20 features based on XGBoost's feature importance for

[3]To compare different feature combinations and reporting significance, we utilize Wilcoxon signed-rank test, which is a non-parametric test for paired data, with signficance level of 0.05.

Features	Time Sessions				
	All	Morning	Noon	Evening	Night
CP	0.556 ± 0.069	0.523 ± 0.140	0.585 ± 0.105	0.558 ± 0.155	0.556 ± 0.127
SS	0.527 ± 0.082	0.519 ± 0.089	0.511 ± 0.098	0.512 ± 0.203	0.548 ± 0.117
SP	**0.678 ± 0.052***	**0.666 ± 0.108**	**0.702 ± 0.109**	**0.668 ± 0.117**	**0.688 ± 0.096**
SS+SP	**0.682 ± 0.053***	**0.677 ± 0.103**	**0.706 ± 0.113**	**0.675 ± 0.120**	**0.680 ± 0.094**
CP+SS	0.552 ± 0.070	0.537 ± 0.117	0.580 ± 0.109	0.529 ± 0.159	0.548 ± 0.128
CP+SS+SP	**0.672 ± 0.056***	**0.659 ± 0.097**	**0.697 ± 0.118**	**0.664 ± 0.136**	**0.671 ± 0.102**

Table 1: AUROC scores at different time sessions for XGBoost model. CP, SS and SP refer to current physiological, sleep statistics and sleep physiological features, respectively. Bold text shows scores with SP feature. '*' represents features with statistical significance over others in 'All' session.

Features	Time Sessions				
	All	Morning	Noon	Evening	Night
CP	0.552 ± 0.061	0.519 ± 0.117	0.573 ± 0.120	0.584 ± 0.167	0.558 ± 0.098
SS	0.501 ± 0.082	0.472 ± 0.125	0.526 ± 0.105	0.489 ± 0.130	0.511 ± 0.136
SP	**0.686 ± 0.049***	**0.666 ± 0.097**	**0.743 ± 0.105**	**0.655 ± 0.189**	**0.688 ± 0.095**
SS+SP	**0.680 ± 0.054***	**0.671 ± 0.115**	**0.724 ± 0.114**	**0.684 ± 0.139**	**0.676 ± 0.102**
CP+SS	0.545 ± 0.067	0.470 ± 0.116	0.577 ± 0.119	0.614 ± 0.148	0.561 ± 0.107
CP+SS+SP	**0.686 ± 0.073***	**0.673 ± 0.091**	**0.714 ± 0.118**	**0.697 ± 0.165**	**0.684 ± 0.121**

Table 2: AUROC scores at different times of the day for multi-modal model. Bold text shows scores with SP feature. '*' indicates represents with statistical significance over others in 'All' session.

the row CP+SS+SP – all of which are sleep physiological features. We report additional metrics of macro F1-scores and per-label F1-scores in the Appendix Tables 5 and 6 for the XGBoost and multimodal network, respectively.

3.2 Effect of Sleep on Stress Detection at Different Times of the Day

As illustrated in Tables 1 and 2, feature combinations with sleep physiological (SP) data achieve the highest mean performance for stress detection at each of the four time-intervals of the day (i.e., morning, afternoon, evening and night), across both the models. While 'noon' achieves achieves the highest AUROC scores, it does not achieve statistical significance[4] over all other sessions, demonstrating consistent influence of sleep features throughout the day.

4 Conclusion and Future Work

In this work, we investigate the role of sleep features in stress detection. Using a month-long dataset of smartwatch-collected physiological signals spanning both daytime and sleep, we extracted HRV-based features and trained two representative models – XGBoost and a multimodal network, to demonstrate the value of incorporating fine-grained sleep information. Our study provides initial evidence for sleep-aware stress prediction, and opens several directions for deeper exploration.

First, given the lack of public datasets that jointly capture sleep and daytime physiological data from wearables, we plan to expand our data collection to a larger and more diverse set of participants. Second, this will enable both improved personalized and generalized stress detection. Specifically, clustering daily physiological data based on sleep features may support cluster-specific personalized modeling [37, 11]. Alternatively, we envision leveraging different *conditioning mechanisms* (such as prefix tuning [22], prompt tuning [33] and FiLM modulation [28]) to integrate sleep information as additional information into pre-trained physiological encoders [26, 6, 7], thereby improving generalization. Overall, we see this line of work as a step towards a more holistic understanding of how sleep influences everyday stress and, more broadly, human affect.

[4]To compare different time sessions and reporting significance, we utilize Mann Whitney U-test, which is a non-parametric test for unpaired data, with a significance level of 0.05.

References

[1] Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[2] Laura SP Bloomfield, Mikaela I Fudolig, Julia Kim, Jordan Llorin, Juniper L Lovato, Ellen W McGinnis, Ryan S McGinnis, Matt Price, Taylor H Ricketts, Peter Sheridan Dodds, et al. Predicting stress in first-year college students using sleep data from wearable devices. *PLOS Digital Health*, 3(4):e0000473, 2024.

[3] 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*, pages 785–794, 2016.

[4] Kayisan M Dalmeida and Giovanni L Masala. Hrv features as viable physiological markers for stress detection using wearable devices. *Sensors*, 21(8):2873, 2021.

[5] Herman J de Vries, Helena JM Pennings, Cees P van der Schans, Robbert Sanderman, Hilbrand KE Oldenhuis, and Wim Kamphuis. Wearable-measured sleep and resting heart rate variability as an outcome of and predictor for subjective stress measures: A multiple n-of-1 observational study. *Sensors*, 23(1):332, 2022.

[6] Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora D Salim, and Akhil Mathur. Crossl: Cross-modal self-supervised learning for time-series through latent masking. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 152–160, 2024.

[7] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–28, 2022.

[8] Ward C Dobbs, Michael V Fedewa, Hayley V MacDonald, Clifton J Holmes, Zackary S Cicone, Daniel J Plews, and Michael R Esco. The accuracy of acquiring heart rate variability from portable devices: a systematic review and meta-analysis. *Sports Medicine*, 49(3):417–435, 2019.

[9] Sunmin Eom, Sunwoo Eom, and Peter Washington. Sim-cnn: Self-supervised individualized multimodal learning for stress prediction on nurses using biosignals. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 155–171. Springer, 2023.

[10] FitRockr. Health data research and healthcare platform for wearables. `https://www.fitrockr.com/research/`, Accessed 2025.

[11] Mikaela Irene Fudolig, Laura SP Bloomfield, Matthew Price, Yoshi M Bird, Johanna E Hidalgo, Julia N Kim, Jordan Llorin, Juniper Lovato, Ellen W McGinnis, Ryan S McGinnis, et al. The two fundamental shapes of sleep heart rate dynamics and their connection to mental health in college students. *Digital Biomarkers*, 8(1):120–131, 2024.

[12] Dana Rose Garfin, Rebecca R Thompson, and E Alison Holman. Acute stress and subsequent health outcomes: A systematic review. *Journal of psychosomatic research*, 112:107–113, 2018.

[13] Shruti Gedam and Sanchita Paul. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9:84045–84066, 2021.

[14] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE transactions on affective computing*, 13(1):440–460, 2019.

[15] Yunjo Han, Panyu Zhang, Minseo Park, and Uichin Lee. Systematic evaluation of personalized deep learning models for affect recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–35, 2024.

[16] Seyedmajid Hosseini, Raju Gottumukkala, Satya Katragadda, Ravi Teja Bhupatiraju, Ziad Ashkar, Christoph W Borst, and Kenneth Cochran. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data*, 9(1):255, 2022.

[17] Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D Salim, Wen Hu, and Aaron J Quigley. Exploring large-scale language models to evaluate eeg-based multimodal data for mental health. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 412–417, 2024.

[18] Salar Jafarlou, Jocelyn Lai, Iman Azimi, Zahra Mousavi, Sina Labbaf, Ramesh C Jain, Nikil Dutt, Jessica L Borelli, and Amir Rahmani. Objective prediction of next-day's affect using multimodal physiological and behavioral data: Algorithm development and validation study. *JMIR Formative Research*, 7(1):e39425, 2023.

[19] Michael Joubert, Jessica Elise Beilharz, Scott Fatt, Yuen Ming Chung, Erin Cvejic, Ute Vollmer-Conna, and Alexander Robert Burton. Stress reactivity, wellbeing and functioning in university students: A role for autonomic activity during sleep. *Stress and Health*, 40(6):e3509, 2024.

[20] Soowon Kang, Woohyeok Choi, Cheul Young Park, Narae Cha, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Heepyung Kim, Yong Jeong, and Uichin Lee. K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels. *Scientific data*, 10(1):351, 2023.

[21] Monika Konjarski, Greg Murray, V Vien Lee, and Melinda L Jackson. Reciprocal relationships between daily sleep and mood: A systematic review of naturalistic prospective studies. *Sleep medicine reviews*, 42:47–58, 2018.

[22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[23] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021.

[24] Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, et al. The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[25] Elizabeth J Mezick, Karen A Matthews, Martica H Hall, J Richard Jennings, and Thomas W Kamarck. Sleep duration and cardiovascular responses to stress in undergraduate men. *Psychophysiology*, 51(1):88–96, 2014.

[26] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, et al. Scaling wearable foundation models. *arXiv preprint arXiv:2410.13638*, 2024.

[27] Mathieu Nollet, William Wisden, and Nicholas P Franks. Sleep deprivation and stress: a reciprocal relationship. *Interface focus*, 10(3):20190092, 2020.

[28] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[29] Julia Reichenberger, Peter Kuppens, Michael Liedlgruber, Frank H Wilhelm, Martin Tiefengrabner, Simon Ginzinger, and Jens Blechert. No haste, more taste: An ema study of the effects of stress, negative and positive emotions on eating behavior. *Biological psychology*, 131:54–62, 2018.

[30] Berrenur Saylam and Özlem Durmaz İncel. Quantifying digital biomarkers for well-being: stress, anxiety, positive and negative affect via wearable devices and their time-based predictions. *Sensors*, 23(21):8987, 2023.

[31] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[32] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine*, 108(3):1287–1301, 2012.

[33] Debaditya Shome, Nasim Montazeri Ghahjaverestan, and Ali Etemad. Naptune: Efficient model tuning for mood classification using previous night's sleep measures along with wearable time-series. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 204–213, 2024.

[34] Maia Ten Brink, Jessica R Dietch, Joshua Tutek, Sooyeon A Suh, James J Gross, and Rachel Manber. Sleep and affect: A conceptual review. *Sleep Medicine Reviews*, 65:101670, 2022.

[35] Fabian Theurl, Michael Schreinlechner, Nikolay Sappler, Michael Toifl, Theresa Dolejsi, Florian Hofer, Celine Massmann, Christian Steinbring, Silvia Komarek, Kurt Mölgg, et al. Smartwatch-derived heart rate variability: a head-to-head comparison with the gold standard in cardiovascular disease. *European Heart Journal-Digital Health*, 4(3):155–164, 2023.

[36] Sofia Triantafillou, Sohrab Saeb, Emily G Lattie, David C Mohr, and Konrad Paul Kording. Relationship between sleep quality and mood: ecological momentary assessment study. *JMIR mental health*, 6(3):e12613, 2019.

[37] 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*, 6(4):1–34, 2023.

[38] Panyu Zhang, Gyuwon Jung, Jumabek Alikhanov, Uzair Ahmed, and Uichin Lee. A reproducible stress prediction pipeline with mobile sensor data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(3):1–35, 2024.

A Related Work

A.1 Stress Prediction on In-The-Wild Datasets

A growing body of work trains and evaluates stress-detection models on *in-the-wild* corpora collected via commercial wearables. We focus on datasets that provide watch-based physiology together with in-situ stress/affect labels: *K-EmoPhone* (smartphone and wearable sensing) [20], a hospital *nurse* dataset enabling continuous monitoring during clinical shifts [16], and *Tesserae* (large-scale workplace sensing with wearables and phones) [24].

Methods built on these datasets illustrate modeling choices under naturalistic variability. On healthcare workers, SIM-CNN pretrains on unlabeled nurse biosignals and fine-tunes with limited labels for stress prediction [9]. In workplace-scale sensing, models trained on *Tesserae* features have been used to predict stress-related outcomes from smartphone and wearable data [30]. For *K-EmoPhone*, the dataset paper reports baseline models using multimodal phone/wearable signals [20], and subsequent work explores personalization and generalization on open, in-the-wild corpora. Notably, Zhang *et al.* [38] perform a comprehensive review of features and techniques for stress detection using *K-EmoPhone* data. They utilize physiological signals (heart rate, skin temperature, skin conductance) as well as auxiliary information (location, phone activity) from the user's *current context*, while evaluating different feature extraction, feature selection and data splitting methods.

In summary, these watch-centric datasets and methods primarily emphasize current-day physiology and context. When sleep is considered, it is typically represented by coarse summaries; detailed nocturnal physiology (e.g., HRV dynamics) is rarely accessible or modeled. Consistent with our introduction, we therefore examine whether incorporating fine-grained overnight sleep physiology, alongside current physiological data, adds value for daily stress detection in naturalistic settings.

A.2 Effect of Sleep on Stress Prediction

Subjective evidence. Daily, within-person studies and systematic reviews suggest that poorer sleep (e.g., lower perceived quality or irregular sleep) tends to precede lower next-day positive affect and

Features	Count
Time-domain HRV: SDNN, RMSSD, SDRMSSD, TINN	4
Frequency-domain HRV: VLF, LF, VHF, LFHF, LFn	5
Non-linear HRV: PIP, PAS, GI, PI, C1d, C2d, ApEn, MSEn, CMSEn, RCMSEn, KFD	11
Other PPG-based: Respiration (breaths/min), Heart Rate (bpm), Spo2	3
Skin Temperature: Temperature (celsius)	1
Accelerometer: Body Battery (%), Motion Intensity, Step Count, Total Count, Zero Crossing Count	5
Total	29

Table 3: Table summarizes all the features derived from the PPG (both HRV and non-HRV features), Temperature and Accelerometer sensors of the Garmin watch.

Hyperparameter	Value	Hyperparameter	Value
`objective`	multi:softprob	`min_child_weight`	5
`booster`	gbtree	`max_depth`	4
`n_estimators`	1000	`colsample_bytree`	0.7
`early_stopping`	20	`alpha`	1
`learning_rate`	0.01	`gamma`	1
`subsample`	0.8	`lambda`	6

Table 4: Hyperparameters for training the XGBoost model.

higher negative affect/stress, and that mood and sleep can influence each other bidirectionally [21, 36, 34]. Population-scale and cohort work also associates prior-night sleep metrics with next-day perceived stress [2].

Objective evidence. From an autonomic perspective, sleep-period electrocardiography (ECG)-derived HRV relates to next-day stress reactivity; shorter sleep has been linked to larger high-frequency HRV withdrawal and slower cardiovascular recovery under laboratory stressors [25]. Field studies monitoring ECG/HRV, subjective stress, and sleep report that same-day or lagged HRV–stress relations may depend on prior sleep quality [5]. While smartwatch photoplethysmography (PPG) can approximate ECG for several global or low-frequency HRV markers, high-frequency, short-term metrics remain more challenging outside controlled settings [35]. Complementing these findings, recent modeling work has begun to integrate previous-night sleep measures with wearable time series for next-day affective-state classification [33].

In summary, literature indicates that sleep—both subjectively and via night physiological activity, bears on next-day stress and affect, yet many wearable pipelines still use coarse aggregates (e.g., duration, resting heart rate, stage fractions). This motivates us to utilize fine-grained sleep physiology for daily stress detection.

B Dataset Details

Here, we provide additional information about the features extracted from the smartwatch, along with the label distribution of the self-reported stress data of our dataset. Table 3 depicts the 29 features derived from the three sensors on the Garmin Venu 3 watch used for the data collection. Amongst all features, the extraction of heart rate variability (HRV) features from sleep physiology, distinguishes our work from other prior efforts towards stress detection.

Figure 2(a) and 2(b) illustrate the distribution of binary stress labels (i.e., low and high) and four time-sessions (i.e., morning, noon, evening and night) across the 28-days of data collection. Overall, our dataset has balanced distribution of stress labels and session splits. The poor representation of data from the first day is due to the lack of sleep data from the night before (this is prior to start of data collection), hence we remove it from our cross validation folds.

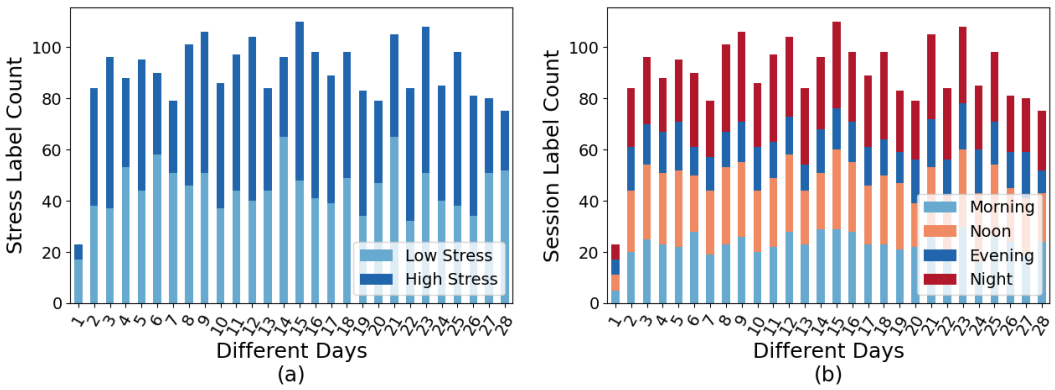

Figure 2: Figure (a) depicts the label distribution and (b) depicts the session distribution across different days in the collected dataset.

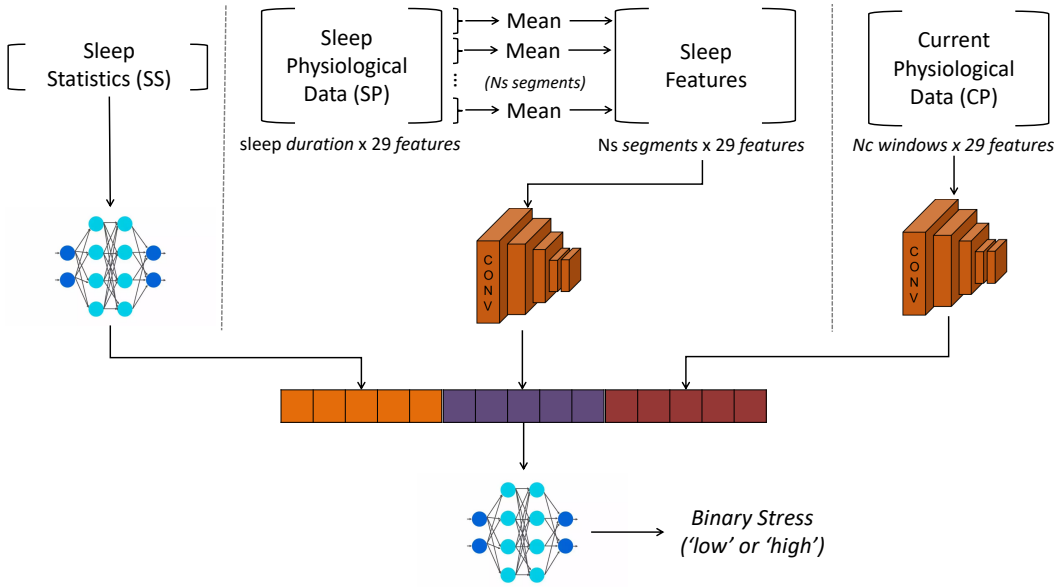

Figure 3: Figure illustrates the architecture of the multi-modal network.

C Training Details and Additional Results

Here, we provide auxiliary information about training the XGBoost and multi-modal network models, as well as additional results. Table 4 enumerates the twelve hyperparameters utilized for training the XGBoost model. Figure 3 depicts the architecture of the multimodal model. The sleep statistics are input to a multi-layer perceptron-based (MLP-based) encoder, while the sleep and current physiological features are input to two different convolutional neural network-based (CNN-based) encoders. Subsequently, all the encoder outputs are concatenated and input to another MLP model which outputs binary stress labels. We train with `batch_size` of 32 samples for 100 `epochs` with a learning rate of 0.0005, and early stopping on validation data with `patience` of 10, and `delta` of 0.0001.

Tables 5 and 6 tabulate the overall stress detection performance of different feature combinations across four different metrics – AUROC, F1-score, and label-specific F1-scores. As highlighted in the tables, the feature combinations with sleep physiological (SP) features achieves improved scores across all metrics. Figure 4 depicts the top-20 features based on their average feature importance scores (across different training folds) based on the *gain* metric. Note that *all* the top-20 features are related to the sleep physiology, highlighting the importance of fine-grained sleep information for stress detection.

Feature	AUROC	macro F1-score	Label=0 F1-score	Label=1 F1-score
CP	0.556 ± 0.069	0.538 ± 0.058	0.524 ± 0.088	0.551 ± 0.089
SS	0.527 ± 0.082	0.523 ± 0.065	0.496 ± 0.097	0.549 ± 0.102
SP	**0.678 ± 0.052**	**0.603 ± 0.059**	**0.603 ± 0.107**	**0.603 ± 0.065**
SS+SP	**0.682 ± 0.053**	**0.607 ± 0.053**	**0.606 ± 0.093**	**0.608 ± 0.055**
CP+SS	0.552 ± 0.070	0.536 ± 0.060	0.511 ± 0.093	0.561 ± 0.092
CP+SS+SP	**0.672 ± 0.056**	**0.606 ± 0.048**	**0.612 ± 0.090**	**0.600 ± 0.055**

Table 5: Performance of different feature combinations using XGBoost model on stress detection. Here, CP: current physiological, SP: sleep physiological, and SS: sleep summary features.

Feature	AUROC	F1-score	Label=0 F1-score	Label=1 F1-score
CP	0.552 ± 0.061	0.524 ± 0.045	0.504 ± 0.086	0.543 ± 0.091
SS	0.501 ± 0.082	0.476 ± 0.053	0.410 ± 0.131	0.542 ± 0.114
SP	**0.686 ± 0.049**	**0.629 ± 0.043**	**0.620 ± 0.081**	**0.638 ± 0.073**
SS+SP	**0.680 ± 0.054**	**0.617 ± 0.060**	**0.599 ± 0.117**	**0.635 ± 0.060**
CP+SS	0.545 ± 0.067	0.525 ± 0.053	0.485 ± 0.100	0.566 ± 0.077
CP+SS+SP	**0.686 ± 0.073**	**0.625 ± 0.073**	**0.611 ± 0.116**	**0.639 ± 0.086**

Table 6: Performance of different feature combinations using multi-modal network on stress detection.

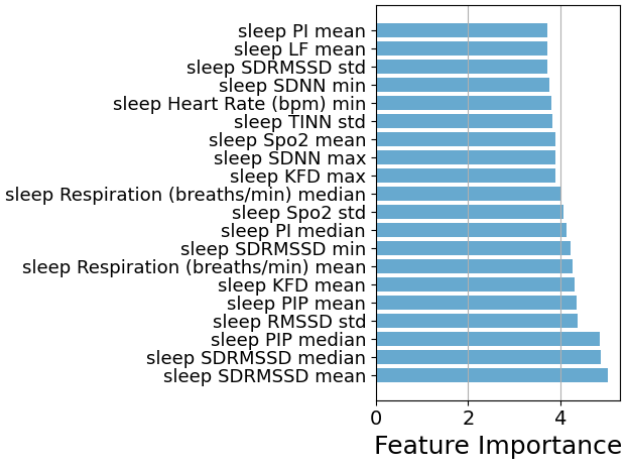

Figure 4: Feature importance scores from XGBoost based on CP+SS+SP feature combination.