This paper is a short version of a longer paper available on arXiv. For the full paper, please see: arxiv.org/abs/2410.20542

PAPAGEI: Open Foundation Models for Optical Physiological Signals

Arvind Pillai[2]*, **Dimitris Spathis**[1,3], **Fahim Kawsar**[1,4], **Mohammad Malekzadeh**[1]
[1]Nokia Bell Labs, UK, [2]Dartmouth College, USA,
[3]University of Cambridge, UK, [4]University of Glasgow, UK

Abstract

Photoplethysmography (PPG) is the most widely used non-invasive technique for monitoring biosignals and cardiovascular health, with applications in both clinical settings and consumer health through wearable devices. However, most models applied to PPG data are task-specific and lack generalizability. Limited previous works often used single-device datasets, did not explore out-of-domain generalization, or did not release their models, hindering open research. Here, we introduce PAPAGEI, the first open foundation model for PPG signals. Pre-trained on more than 57,000 hours of 20 million unlabeled PPG signals using publicly available datasets exclusively, PAPAGEI is evaluated against popular time-series foundation models and other benchmarks on 18 diverse tasks spanning cardiovascular health, sleep disorders, pregnancy monitoring, and wellbeing assessment. PAPAGEI's architecture incorporates a novel representation learning approach that examines differences in PPG signal morphology across individuals, enabling it to capture rich representations. Across 18 clinically-relevant classification and regression tasks, PAPAGEI outperforms baselines in 13, resulting in an average improvement of 6.3% and 2.9%, respectively. Notably, it can be used out of the box as both a feature extractor and an encoder for other multimodal models, opening up new opportunities for multimodal health monitoring [2].

1 Introduction

Photoplethysmography (PPG) is a technique that enables non-invasive monitoring of physiological signals by capturing changes in blood flow volume through light-based (optical) sensing [1]. In hospitals and clinics, PPG is used for monitoring blood oxygen and heart rate, and has also been integrated into consumer devices like smartwatches, making continuous health monitoring accessible in both clinical settings and daily life. This dual role emphasizes PPG's importance in acute medical care and long-term health management. PPG signals have demonstrated utility in tracking various conditions, from cardiovascular health to mood and sleep disorders [2, 3, 4, 5, 6, 7, 8]. However, PPG data presents significant challenges for machine learning (ML), including difficulties in data annotation, susceptibility to noise and motion artifacts [9], and inherent variability due to factors like skin tone and body composition [10]. These challenges have resulted in small, task-specific datasets that limit the development of robust, generalizable models.

*Work done while at Nokia Bell Labs.
[2]Code available at `https://github.com/Nokia-Bell-Labs/papagei-foundation-model`

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

Therefore, we introduce **PAPAGEI**, a robust set of pre-trained models designed as a backbone for diverse PPG tasks. Our **contributions** include **large-scale pre-training** on 57,000 hours of public PPG data from sources like VitalDB [11], MIMIC-III [12], and MESA[13]. To enhance pre-training, we develop a **novel PPG-aware self-supervised learning** framework and perform **comprehensive evaluations** across 18 clinically relevant tasks—spanning cardiovascular health, sleep disorders, pregnancy monitoring, and overall well-being. By releasing our models and code, PAPAGEI provides a strong foundation for applying large-scale models to domain-specific biosignals and promotes future research in this field.

2 Related work

Self-supervised learning (SSL) allows learning general representations from unlabeled datasets, with promising applications in physiological signal analysis [14, 15, 16, 17, 18, 19, 20, 21], showing the potential of PPG embeddings for various health outcomes and applications [22, 23, 24, 25]. For example, REGLE [23] showed that embedding PPG signals can improve genetic discovery and risk prediction outcomes. However, the field lacks widely available pre-trained models for PPG data, with existing studies often being limited by single-device datasets, lack of out-of-domain generalization, or unavailability of released models. While generic time series foundation models like Chronos [26] and Moment [27] have gained traction, they often lack significant physiological data representation. Our work takes a domain-specific approach, focusing exclusively on PPG data to capture its unique characteristics and complexities, building upon the growing interest in modality-specific foundation models such as those for ECG [28, 29] and EEG [30].

3 Model

We use self-supervised learning to train PAPAGEI's deep neural network encoder. First, we propose PAPAGEI-P, a patient contrastive approach to maximize agreement between signals from the same participant. While [22] implements a similar strategy, they do not evaluate on public datasets. Importantly, we propose PAPAGEI-S, a morphology-aware model that maximizes agreement between signals that exhibit similar morphology. Formally, given a dataset $\mathcal{D} = \{\mathbf{p}^1, \mathbf{p}^2, \cdots, \mathbf{p}^S\}$ representing PPG signals from S subjects, a PPG signal $\mathbf{p}^s \in \mathbb{R}^n$ is defined as a time series that models the changes in light intensity due to arterial blood flow. To model granular changes in PPG signal obtained from a subject s, we segment $\mathbf{p}^s$ without overlap to obtain $X^s = \{\mathbf{x}_1^s, \mathbf{x}_2^s, \cdots \mathbf{x}_{\lfloor \frac{n}{l} \rfloor}^s\}$. Here, the length l of each segment $\mathbf{x}$ is a product of the sampling frequency and the desired time window.

Table 1: PAPAGEI's pre-training datasets.

Dataset	#People	#Segments	Hours
VitalDB	5,866	6,248,100	17,355
MIMIC-III	5,596	7,196,401	19,990
MESA	2,055	7,306,705	20,296
Total	13,517	20,751,206	57,641

Participant-aware objective. We define a positive pair as any two distinct segments from the same subject, denoted as $\{(\mathbf{x}_i^s, \mathbf{x}_j^s) | i \neq j\}$. Next, we apply a series of random time series augmentations such as random cropping, adding Gaussian noise, time flipping, negation, and magnitude scaling [31]. During training, the two randomly sampled positive pairs are passed through the encoder E and projection P to obtain embeddings denoted $(\mathbf{z}_i^s, \mathbf{z}_j^s)$. Given a batch of embeddings from N distinct subjects with positive pairs of the form $(\mathbf{z}_i^s, \mathbf{z}_j^s)$, the model optimizes the NT-Xent loss given by: $\mathcal{L}_p = \frac{1}{2}(\ell_p(i,j) + \ell_p(j,i))$, where $\ell_p(i,j) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(sim(\mathbf{z}_i^s, \mathbf{z}_i^s)/\tau)}{\sum_{s \neq r} \exp(sim(\mathbf{z}_i^s, \mathbf{z}_j^r)/\tau)}$. Contrastingly, vanilla SimCLR [32] uses positive pairs as augmented versions of randomly sampled PPG segments.

Segment-aware objective. To incorporate morphology into self-supervised learning, we introduce a morphology augmentation module prior to training that computes three PPG metrics: (1) Stress-induced Vascular Response Index (sVRI) [33, 34]: the ratio of mean PPG signal between post- to pre-systolic phases, (2) Inflection Point Area ratio (IPA) [35]: the ratio of systolic to diastolic areas defined by the dicrotic notch, and (3) Signal quality index (SQI): skewness of the signal as an indicator of quality [36]. These metrics complement each other, with sVRI capturing amplitude variations, IPA reflecting signal width, and SQI addressing cases where IPA cannot be computed due to poor-quality signals lacking a dicrotic notch.

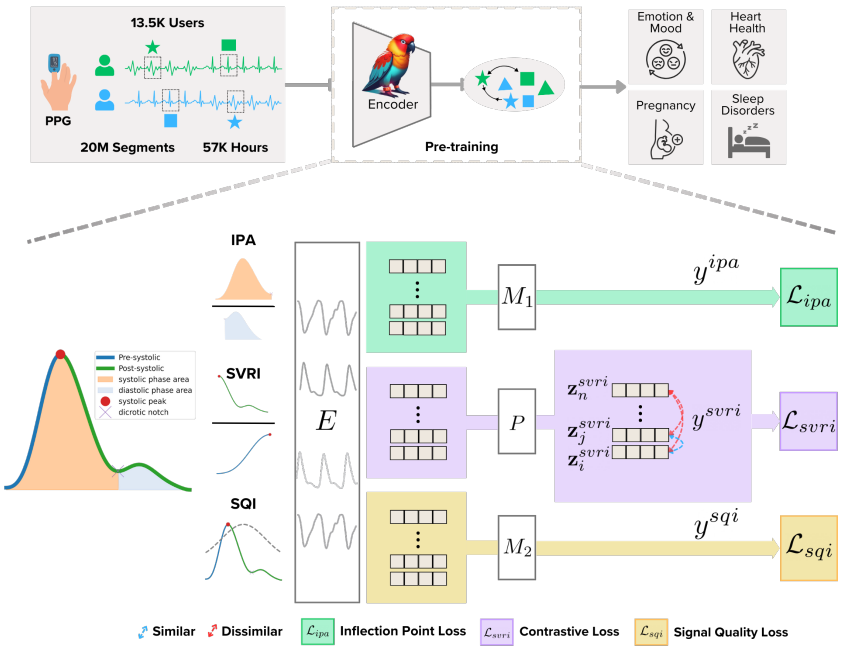

Figure 1: PAPAGEI Overview: (Top) We train a large PPG model using public data from 13.5K users comprising 20M segments for downstream medical tasks. (Bottom) PAPAGEI-S: (Left) We calculate morphology metrics (IPA, SVRI, and SQI) for each PPG segment. (Middle) A batch of PPG signals, with morphology, is passed to an encoder (E) to extract embeddings. (Right) The projection head (P) contrasts signals based on SVRI, while expert heads (M_1 and M_2) refine embeddings by predicting IPA and SQI values, respectively.

$$\ell_s(i,j) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(sim(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k\neq j} \exp\left(sim(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{1}$$

$$\mathcal{L}_{svri} = \frac{1}{2}(l_s(i,j) + l_s(j,i)) \quad \mathcal{L}_{ipa} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{ipa} - \hat{y}_i^{ipa}\right| \quad \mathcal{L}_{sqi} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{sqi} - \hat{y}_i^{sqi}\right| \tag{2}$$

$$\mathcal{L}_s = \alpha\mathcal{L}_{svri} + (1-\alpha)\left(\mathcal{L}_{ipa} + \mathcal{L}_{sqi}\right), \text{where } \alpha \in [0,1] \tag{3}$$

The morphology augmentation module takes an input time series $\mathbf{x}$ and outputs $y = \{y^{svri}, y^{ipa}, y^{sqi}\} \in \mathbb{R}^3$ (Figure 1 middle). To compute positive pairs, we first discretize y^{svri} into a predefined set of $n = 8$ bins, where $y^{svri} \in \{0, 1, \ldots, n\}$. We define positive pairs based on the sVRI labels as $\{(\mathbf{x}_i, \mathbf{x}_j) | y_i^{svri} = y_j^{svri}, i \neq j\}$. In PAPAGEI-S, given a batch of N PPG signals and their morphology, we optimize three heads. First, we extract the embeddings $Z = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$ from the projection P (Figure 1 right middle), and compute contrastive loss for sVRI (equation 1). Next, we use the embeddings $H = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_N\}$ to predict the IPA ($\hat{\mathbf{y}}^{ipa}$) and SQI ($\hat{\mathbf{y}}^{sqi}$) using the mixture of expert heads M_1 and M_2. These heads are optimized using the mean absolute error (equation 2). Finally, the overall PAPAGEI-S training objective given in equation 3 ($\alpha = 0.6$).

4 Experiments

Training. We pre-train PAPAGEI using a ResNet-style CNN with 5/5.7M parameters on three large public datasets: VitalDB [11], MIMIC-III [12], and the MESA sleep sub-study [13, 37]. After data curation, we apply several pre-processing steps: (1) a 4th-order Chebyshev bandpass filter (0.5–12Hz) [38, 39], (2) segmentation into 10-second windows [40, 41], (3) removal of segments with over 25% flatline data [42], (4) Z-score normalization [22], and (5) downsampling to 125Hz. This results in

Table 2: Downstream tasks comparison against pre-trained models (parameter size and references are denoted next to their names). 95% CIs are reported in square brackets and the best value is **bolded**.

Classification (AUROC ↑)	**REGLE** [23] (0.07M)	**Chronos** [26] (200M)	**Moment** [27] (385M)	**PAPAGEI-P** (5M)	**PAPAGEI-S** (5.7M)
ICU Admission	0.57 [0.52-0.62]	0.73 [0.68-0.80]	0.72 [0.70-0.80]	0.73 [0.67-0.78]	**0.79** [0.75-0.82]
Smoker	0.54 [0.47-0.59]	0.62 [0.57-0.67]	0.62 [0.56-0.67]	**0.64** [0.58-0.69]	0.61 [0.56-0.66]
Mortality	0.55 [0.52-0.59]	**0.68** [0.65-0.71]	0.67 [0.63-0.71]	0.67 [0.63-0.71]	0.67 [0.63-0.70]
Sleep-disordered Breathing	0.45 [0.30-0.61]	0.58 [0.35-0.82]	0.45 [0.23-0.66]	0.54 [0.23-0.66]	**0.70** [0.57-0.84]
Hypertension	0.47 [0.34-0.58]	0.57 [0.43-0.71]	0.75 [0.64-0.85]	0.74 [0.55-0.90]	**0.77** [0.68-0.87]
Valence	0.55 [0.52-0.57]	0.56 [0.53-0.59]	**0.57** [0.54-0.59]	0.53 [0.51-0.56]	0.56 [0.54-0.59]
Arousal	0.51 [0.52-0.58]	0.57 [0.54-0.60]	0.56 [0.53-0.58]	**0.58** [0.55-0.61]	0.55 [0.52-0.57]
Mood Disturbance	0.41 [0.16-0.66]	0.43 [0.21-0.68]	0.55 [0.33-0.78]	0.53 [0.27-0.78]	**0.56** [0.33-0.77]
Pregnancy stage	0.64 [0.57-0.63]	**0.81** [0.79-0.82]	0.76 [0.74-0.78]	0.74 [0.72-0.76]	0.78 [0.75-0.80]
Average	0.52 ± 0.06	0.62 ± 0.10	0.63 ± 0.09	0.63 ± 0.08	**0.67 ± 0.09**
Regression (MAE ↓)					
AHI > 3%	15.54 [14.20-16.69]	14.06 [13.05-15.16]	14.23 [13.04-15.42]	13.85 [12.43-15.49]	**12.97** [11.87-14.05]
AHI > 4%	12.64 [11.47-13.78]	11.57 [10.51-12.72]	11.80 [10.79-12.93]	11.24 [9.71-12.87]	**10.56** [9.59-11.62]
Systolic BP (PPG-BP)	16.32 [13.87-19.13]	16.91 [13.31-19.34]	14.50 [11.98-17.31]	**13.60** [10.65-16.51]	14.39 [12.53-16.45]
Diastolic BP (PPG-BP)	9.30 [7.94-10.87]	10.26 [8.13-12.57]	9.53 [8.28-10.96]	8.88 [7.33-10.76]	**8.71** [7.18-10.01]
Average HR	6.88 [5.81-8.12]	8.51 [7.05-10.07]	4.41 [3.48-5.48]	**3.47** [2.74-4.32]	4.00 [3.34-4.67]
HR	16.35 [16.20-16.50]	9.65 [9.50-9.79]	**8.82** [8.68-8.96]	10.92 [10.80-11.04]	11.53 [11.40-11.66]
Gestation Age	7.28 [7.16-7.39]	**5.69** [5.54-5.85]	6.24 [6.10-6.37]	6.40 [6.21-6.59]	6.05 [5.91-6.17]
Systolic BP (VV)	15.88 [13.67-18.36]	17.24 [14.57-20.13]	14.71 [12.38-17.29]	19.11 [16.26-22.23]	**14.65** [12.50-16.78]
Diastolic BP (VV)	8.65 [7.16-10.27]	10.53 [8.91-12.19]	10.53 [8.91-12.19]	10.87 [9.10-12.98]	**8.29** [6.61-10.22]
Average	12.09 ± 3.83	11.60 ± 3.60	10.43 ±3.46	10.92 ± 4.25	**10.12 ± 3.47**

20M segments from 13.5K people (Table 1). We train PAPAGEI for 15,000 steps on 8 V100 GPUs with a learning rate of 10^{-4}.

Linear Probing. For downstream evaluation, we use the following datasets and tasks (Appendix A): (1) VitalDB [11]: ICU admission, (2) MESA [13, 37]: AHI > 3% and 4% oxygen desaturation, and smoking status, (3) MIMIC-III [12]: mortality, (4) SDB [43]: sleep-disordered breathing, (5) PPG-BP [6]: systolic/diastolic blood pressure, heart rate, and hypertension, (6) WESAD [44]: valence and arousal, (7) PPG-DaLia [5]: heart rate, (8) ECSMP [45]: mood disturbance, (9) nuMoM2B [46, 13]: pregnancy stage and gestation age, and (10) VitalVideos [47]: systolic/diastolic blood pressure. We split datasets into training/validation/testing sets (80/10/10 for ID and 60/20/20 for OOD), ensuring no participant overlap. Feature embeddings from the projection layer are extracted for downstream prediction. For binary classification, we use logistic regression with ROC-AUC, whereas regression tasks use ridge regression with mean absolute error (MAE). Further, we compute 95% confidence intervals using bootstrapping (500 sampling runs with replacement). We compare our model to pre-trained baselines REGLE [23], Chronos [26], and Moment [27].

5 Results & Discussion

From Table 3, we observe that PAPAGEI surpasses the baseline models in 13 out of 18 tasks. Specifically, PAPAGEI-S achieves improvements of 2.6%, 1.9%, 0.4%, and 4.1% in hypertension, diastolic BP (PPG-BP), systolic BP, and diastolic BP, respectively. These findings highlight the effectiveness of our approach, especially considering that changes in blood pressure affect blood volume, which is captured by sVRI. Notably, our model with 5.7 million parameters, is more efficient than other time series foundation models, largely because CNN architectures are particularly well-suited for PPG signals [22, 24]. From a practical standpoint, having models that can be deployed on mobile phones, finger oximeters, and wearables—the main sources of PPG signals—is advantageous. While REGLE was trained on the UK Biobank, its small genetics-focused model struggles to capture generalizable features, whereas, among the larger models, Moment outperforms Chronos due to its training on a subset of physiological data. Additionally, our preliminary ablation studies show that the combined model (sVRI + IPA + SQI) delivers the best performance, with the sVRI contrastive component contributing the most. Similarly, early comparisons with contrastive baselines such as SimCLR and TF-C demonstrate that our method outperforms them in 12 out of 18 tasks. In the future, we plan to explore model scaling, alternative backbone architectures like CNN-Transformer, and different input sampling rates. Overall, PAPAGEI-S's morphology-aware training proves effective across a variety of classification and regression tasks on diverse datasets.

References

[1] Peter H Charlton, John Allen, Raquel Bailón, Stephanie Baker, Joachim A Behar, Fei Chen, Gari D Clifford, David A Clifton, Harry J Davies, Cheng Ding, et al. The 2023 wearable photoplethysmography roadmap. *Physiological measurement*, 44(11):111001, 2023.

[2] Tariq Sadad, Syed Ahmad Chan Bukhari, Asim Munir, Anwar Ghani, Ahmed M El-Sherbeeny, and Hafiz Tayyab Rauf. Detection of cardiovascular disease based on ppg signals using machine learning with cloud computing. *Computational Intelligence and Neuroscience*, 2022(1):1672677, 2022.

[3] Arrozaq Ave, Hamdan Fauzan, S Rhandy Adhitya, and Hasballah Zakaria. Early detection of cardiovascular disease with photoplethysmogram (ppg) sensor. In *2015 international conference on electrical engineering and informatics (ICEEI)*, pages 676–681. IEEE, 2015.

[4] Andriy Temko. Accurate heart rate monitoring during physical exercises using ppg. *IEEE Transactions on Biomedical Engineering*, 64(9):2016–2024, 2017.

[5] Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[6] Yongbo Liang, Zhencheng Chen, Guiyong Liu, and Mohamed Elgendi. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in china. *Scientific data*, 5(1):1–7, 2018.

[7] Serj Haddad, Assim Boukhayma, and Antonino Caizzone. Continuous ppg-based blood pressure monitoring using multi-linear regression. *IEEE journal of biomedical and health informatics*, 26(5):2096–2105, 2021.

[8] Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of deep learning based blood pressure prediction from ppg and rppg signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3820–3830, 2021.

[9] Amir Hosein Afandizadeh Zargari, Seyed Amir Hossein Aqajari, Hadi Khodabandeh, Amir Rahmani, and Fadi Kurdahi. An accurate non-accelerometer-based ppg motion artifact removal technique using cyclegan. *ACM Transactions on Computing for Healthcare*, 4(1):1–14, 2023.

[10] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):18, 2020.

[11] Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[13] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

[14] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*, 2021.

[15] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

[16] Hugh Chen, Scott M Lundberg, Gabriel Erion, Jerry H Kim, and Su-In Lee. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. *NPJ Digital Medicine*, 4(1):167, 2021.

[17] Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pages 11964–11974. PMLR, 2021.

[18] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 69–78, 2021.

[19] Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

[20] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[21] Pritam Sarkar and Ali Etemad. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554, 2020.

[22] Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

[23] Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, et al. Unsupervised representation learning on high-dimensional clinical data improves genomic discovery and prediction. *Nature Genetics*, pages 1–10, 2024.

[24] Wei-Hung Weng, Sebastien Baur, Mayank Daswani, Christina Chen, Lauren Harrell, Sujay Kakarmath, Mariam Jabara, Babak Behsaz, Cory Y McLean, Yossi Matias, et al. Predicting cardiovascular disease risk using photoplethysmography and deep learning. *PLOS Global Public Health*, 4(6):e0003204, 2024.

[25] Cheng Ding, Zhicheng Guo, Zhaoliang Chen, Randall J Lee, Cynthia Rudin, and Xiao Hu. Siamquality: a convnet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8):085004, 2024.

[26] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[27] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[28] Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024.

[29] Junho Song, Jong-Hwan Jang, Byeong Tak Lee, DongGyun Hong, Joon-myoung Kwon, and Yong-Yeon Jo. Foundation models for electrocardiograms. *arXiv preprint arXiv:2407.07110*, 2024.

[30] Zhizhang Yuan, Daoze Zhang, Junru Chen, Geifei Gu, and Yang Yang. Brant-2: Foundation model for brain signals. *arXiv preprint arXiv:2402.10251*, 2024.

[31] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, and Cecilia Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[32] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pages 1597–1607, 2020.

[33] Yongqiang Lyu, Xiaomin Luo, Jun Zhou, Chun Yu, Congcong Miao, Tong Wang, Yuanchun Shi, and Ken-ichi Kameyama. Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 857–866, 2015.

[34] Xiao Zhang, Yongqiang Lyu, Tong Qu, Pengfei Qiu, Xiaomin Luo, Jingyu Zhang, Shunjie Fan, and Yuanchun Shi. Photoplethysmogram-based cognitive load assessment using multi-feature fusion model. *ACM Transactions on Applied Perception (TAP)*, 16(4):1–17, 2019.

[35] L Wang, Emma Pickwell-MacPherson, YP Liang, and Yuan Ting Zhang. Noninvasive cardiac output estimation using a novel photoplethysmogram index. In *2009 annual international conference of the IEEE engineering in medicine and biology society*, pages 1746–1749. IEEE, 2009.

[36] Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016.

[37] Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6):877–888, 2015.

[38] Denis G Lapitan, Dmitry A Rogatkin, Elizaveta A Molchanova, and Andrey P Tarasov. Estimation of phase distortions of the photoplethysmographic signal in digital iir filtering. *Scientific Reports*, 14(1):6546, 2024.

[39] Yongbo Liang, Mohamed Elgendi, Zhencheng Chen, and Rabab Ward. An optimal filter for short photoplethysmogram signals. *Scientific data*, 5(1):1–12, 2018.

[40] Christina Orphanidou. Quality assessment for the photoplethysmogram (ppg). *Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations*, pages 41–63, 2018.

[41] Bojana Koteska, Ana Madevska Bodanova, Hristina Mitrova, Marija Sidorenko, and Fedor Lehocki. A deep learning approach to estimate spo2 from ppg signals. In *Proceedings of the 9th International Conference on Bioinformatics Research and Applications*, pages 142–148, 2022.

[42] BioBSS Documentation. Biobss: Biosignal processing toolbox. `https://biobss.readthedocs.io/en/latest/`, 2023. Accessed: 2024-09-10.

[43] Ainara Garde, Parastoo Dehkordi, Walter Karlen, David Wensley, J Mark Ansermino, and Guy A Dumont. Development of a screening tool for sleep disordered breathing in children using the phone oximeter™. *PloS one*, 9(11):e112959, 2014.

[44] 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.

[45] Zhilin Gao, Xingran Cui, Wang Wan, Wenming Zheng, and Zhongze Gu. Ecsmp: A dataset on emotion, cognition, sleep, and multi-model physiological signals. *Data in Brief*, 39:107660, 2021.

[46] Francesca L Facco, Corette B Parker, Uma M Reddy, Robert M Silver, Judette M Louis, Robert C Basner, Judith H Chung, Frank P Schubert, Grace W Pien, Susan Redline, et al. Numom2b sleep-disordered breathing study: objectives and methods. *American journal of obstetrics and gynecology*, 212(4):542–e1, 2015.

[47] Pieter-Jan Toye. Vital videos: A dataset of videos with ppg and blood pressure ground truths. *arXiv preprint arXiv:2306.11891*, 2023.

APPENDIX

A Downstream Tasks

Table 3: The task evaluation benchmark of PAPAGEI. Datasets highlighted in grey are unseen during training, thus, the corresponding tasks are out-of-domain. The rest were used for pre-training but their test sets and labels are held out. For task type, B/M/R refers to Binary classification, Multi-class classification (#classes), and Regression, respectively.

#ID	Dataset	SR (Hz)	Collected by	Task	Task Type	#Subjects (#Samples)
T1	VitalDB [11]	500	ICU monitor	ICU admission (Yes/No)	B	5866
T2				Operation Type	M (11)	5866
T3	MESA [37]	256	Polysomnography finger	Smoker	B	2055
T4				AHI $>$ 3% Oxygen Desat.	R	2055
T5				AHI $>$ 4% Oxygen Desat.	R	2055
T6	MIMIC-III [12]	125	ICU Monitor	Mortality	B	5596
T7	SDB [43]	62.5	Finger Pulse Ox	Sleep-Disordered Breathing	B	146
T8	PPG-BP [6]	1000	Finger Pulse Ox	Systolic Blood Pressure	R	219
T9				Diastolic Blood Pressure	R	219
T10				Average Heart Rate	R	219
T11				Hypertension	B	219
T12	WESAD [44]	64	Wrist	Valence	B	15 (4497)
T13				Arousal	B	15 (4497)
T14	PPG-DaLiA [5]	64	Wrist	Heart Rate	R	15 (64697)
T15				Activity	M (9)	15 (64697)
T16	ECSMP [45]	64	Wrist	Mood Disturbance	B	89
T17	nuMom2B [46]	75	Polysomnography finger	Pregnancy stage (early/late)	B	3163 (5337)
T18				Gestation Age	R	3163 (5337)
T19	VV (Skin Tone) [47]	60	Finger	Systolic BP	R	231
T20				Diastolic BP	R	231