Towards Characterizing Knowledge Distillation of PPG Heart Rate Estimation Models

Kanav Arora
University of Washington
Seattle, WA
kanava@cs.washington.edu

Girish Narayanswamy
University of Washington
Seattle, WA
girishvn@uw.edu

Shwetak Patel
University of Washington
Seattle, WA
shwetak@cs.washington.edu

Richard Li
University of Washington
Seattle, WA
lichard@cs.washington.edu

Abstract

Heart rate estimation from photoplethysmography (PPG) signals generated by wearable devices such as smartwatches and fitness trackers has significant implications for the health and well-being of individuals. Although prior work has demonstrated deep learning models with strong performance in the heart rate estimation task, in order to deploy these models on wearable devices, these models must also adhere to strict memory and latency constraints. In this work, we explore and characterize how large pre-trained PPG models may be distilled to smaller models appropriate for real-time inference on the edge. We evaluate four distillation strategies through comprehensive sweeps of teacher and student model capacities: (1) *hard distillation*, (2) *soft distillation*, (3) *decoupled knowledge distillation (DKD)*, and (4) *feature distillation*. We present a characterization of the resulting scaling laws describing the relationship between model size and performance. This early investigation lays the groundwork for practical and predictable methods for building edge-deployable models for physiological sensing.

1 Introduction

Wearable devices such as smartwatches and fitness trackers have enabled the collection of in-situ datasets of sensor signals with the potential to support individuals in tracking and monitoring their health and well-being. Amongst other signals, photoplethysmography (PPG), a method for optical estimation of blood volume pulse (BVP), has shown utility in allowing individuals to gauge their cardiovascular health [11, 16]. The growing ubiquity of wearable devices has led to the accumulation of large PPG datasets [7, 9, 12] and the subsequent training of large neural models useful in estimating cardiac function such as heart rate and heart rate variability [8, 11, 14, 18, 19]. These developments represent significant progress towards end-user applications, such as providing real-time feedback in exercise contexts (e.g., heart rate response to exercise intensity) as well as passive screening of diseases (e.g., hypertension).

Despite the success of these large models across a variety of sensor data tasks, their significant computational requirements pose a barrier to adoption and limit their utility [6]. While edge models such as those running on wearables better preserve privacy and better support real-time feedback, large sensor models may struggle to realize these gains. More work is thus needed to develop and characterize methods for enabling large physiological sensing models to effectively scale to the edge.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

Table 1: Experimental variables for characterizing the process of distilling PPG models.

Name	Description	Values
Strategy	Procedure for distillation	Hard Distillation, Soft Distillation, Decoupled Knowledge Distillation, Feature Distillation
Teacher size	# of residual blocks (# of parameters) in teacher model	2 (33,724), 3 (44,156), 4 (54,588), 5 (97,852), 6 (139,196), 8 (221,884), 10 (534,460), 12 (863,676)
Student size	# of residual blocks (# of parameters) in student model	1 (23,292), 2 (33,724), 3 (44,156), 4 (54,588), 5 (97,852), 6 (139,196), 8 (221,884), 10 (534,460)

Prior work has established the utility of knowledge distillation [2, 4], where efficient student models learn from larger, high capacity, pretrained teacher models. For example, DistilBERT [15] has found success in optimizing language models for edge deployments while retaining strong performance. More similar to wearable physiological sensing, prior work has found success in distilling audio [10] and accelerometer models [17] useful for human activity recognition. However, while knowledge distillation has been established as a powerful tool in developing compute-efficient models, there has been little exploration into the characterization of these methods, making it difficult to predict the performance of a distilled model. Only recently have scaling laws that govern the distillation of language models been established to predictably compute distilled language model performance [1].

Building off these ideas, our work takes a first step towards establishing predictable distillation performance in the domain of physiological sensing. Specifically, for the task of PPG heart rate estimation, we evaluate four distillation strategies across different student and teacher model capacities and characterize the effect of these variables on distilled model size. We further compare the interplay between model computational requirements (i.e., memory consumption and inference time) and distilled performance. We confirm that distilled models improve upon models trained from scratch, find that decoupled knowledge distillation outperforms other evaluated strategies, demonstrate that the performance of distilled models follow a characterization exponential scaling curve, and observe that these scaling behaviors vary by model architecture.

2 Methods

Distillation Experiment Setup. We characterize the distillation scaling behavior of physiological sensing models for PPG across a number of teacher and student model sizes. Specifically, we employ the variant of a 1D-ResNet backbone [3] used by Meier et al. [7] to classify the instantaneous heart rate given a PPG signal window. To vary model capacity, we sweep across the number of residual blocks (i.e., resulting in an approximately exponential sweep of model parameters) for student and teacher models, as illustrated in Table 1. We further explore the following four distillation strategies:

- *Hard Distillation:* The teacher model's predictions (i.e., the final $argmax$ output) are used as labels for training the student model, helping it mimic the discrete decision boundaries of the teacher.
- *Soft Distillation:* The student model is trained on the output probability distributions of the teacher model, encoding richer information about inter-class relationships and uncertainty [4].
- *Decoupled Knowledge Distillation (DKD):* The teacher model's outputs are separated into target class and non-target class distillation components in the student model's loss to introduce flexibility in weighting the significance of true label and incorrect label probabilities [20].
- *Feature Distillation:* Moving beyond operating on model outputs, in *feature distillation*, the student model is trained to match the learned feature maps of the teacher model, aligning their intermediate representation spaces [13].

Heart rate estimation performance is evaluated via Mean Absolute Error (MAE) in beats per minute (BPM). The performance of all distilled student models are evaluated against a corresponding model of the same size trained from scratch.

Training Procedure. All models were trained for 300 epochs at a learning rate of $5 * 10^{-4}$ using a cross-entropy loss. Following the task formulation given by Meier et al. [7], all models are trained

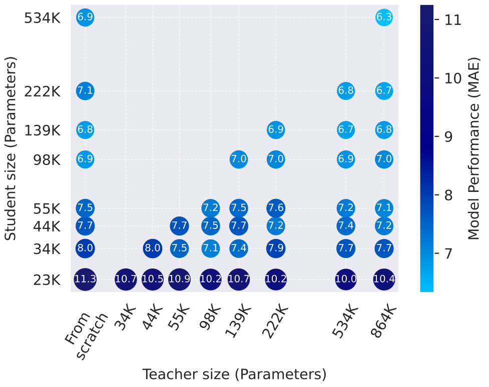

Figure 1: ***DKD* Distilled Model Performance Across Different Student / Teacher Capacities.** Color and size both encode MAE metric for instantaneous heart-rate prediction. The “From scratch” column denotes baseline models trained from scratch rather than distilled from a teacher.

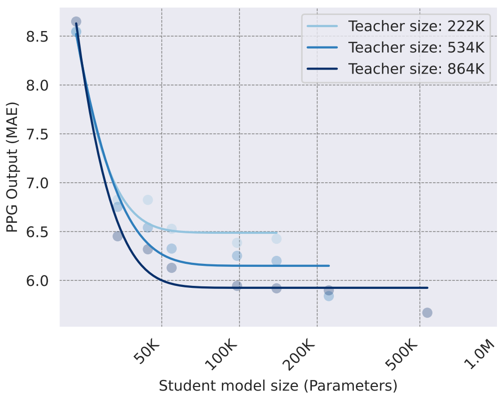

Figure 2: ***DKD* Distilled Model Scaling Behavior.** Scaling curves for distilled student model performance as a function of student and teacher model size. Note that experimental conditions with smaller teacher sizes yielded too few data points to effectively fit a curve.

to predict the instantaneous heart rate via classification by making a decision between 180 classes corresponding to heart rate values between 30 to 210 BPM.

Datasets. For all experiments, we leverage three free-living PPG datasets containing a total of 107 hours of PPG sensor signals: (1) WildPPG [7], (2) PPG-DaLiA [12], (3) GalaxyPPG [9]. Following prior work, we use only the green channel of the PPG sensor, resampled to 25 Hz and segmented into 8-second windows with 2-second strides [7, 12]. Each dataset includes heart rate ground truth (i.e., in BPM) derived via an electrocardiogram (ECG) signal. We generate participant-independent train-test splits by taking data from 80% of the participants for training, and data from 20% of the participants for evaluation. We conduct 2-fold cross validation across all experiments.

3 Results

Distilled models outperform those trained from scratch. In Figure 1, we show the results of our distillation experiment using the *DKD* strategy. The left-most column, “From scratch”, denotes baseline models of a given size trained from scratch rather than distillation. We find that this baseline is consistent with prior work (i.e., the target model size with 8-blocks yields a similar MAE to the results reported by Meier et al. [7] using the same model). In general, we observe that smaller models exhibit worse MAE performance, and that distillation always improves performance over training from scratch. We note that larger teacher models generally exhibit improved performance, and hypothesize that too-large models may overfit easily, resulting in degradation. These results support that gains can be obtained in terms of performance from model distillation.

***DKD* outperforms other strategies.** As shown in Table 2, we find that *DKD* performs the best of the four strategies evaluated across all model size configurations, including across different teacher sizes not shown in the table. *DKD* is followed in performance by *feature distillation*, then *soft distillation*, and finally *hard distillation*. Out of the logit-based strategies, *hard distillation* performed the worst due to the lack of information encoded in its discrete labels, and *soft distillation* performed marginally better. *DKD*, on the other hand, poses the clearest advantage in being able to flexibly weigh true label and incorrect label probabilities, particularly in our task framing where the classification bins are semantically ordinal. Through hyperparameter search, we found $\alpha = 1$, $\beta = 8$, temperature $\tau = 2$, cross-entropy loss weight $CE = 1$ to work best. With these parameters, non-target class distillation (NCKD) probabilities are weighed 8 times more than target class distillation (TCKD) probabilities. Our results indicate that while our small models may not have the capacity to learn a rich representation when trained from scratch. Instead, by regressing to richer probability labels (via distillation) rather than the original BPM ground truth labels, the student models are able to

Table 2: Model performance (MAE) across different distillation strategies and student model sizes. Teacher model size is fixed at 12 blocks.

Distillation	**Student Model Size (Blocks)**							
Strategy	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**10**
Hard	11.734	10.418	9.256	7.478	7.208	6.983	6.830	6.493
Soft	10.380	7.703	7.200	7.111	7.042	6.801	6.679	6.327
DKD	**8.899**	**6.772**	**6.689**	**6.849**	**6.522**	**6.291**	**5.959**	**5.759**
Feature	9.397	7.200	6.952	6.914	6.872	6.800	6.659	6.409

more closely mimic and learn the internal representation of the larger teacher models, leading to improved performance. We thus show that our student models are small enough to learn a strong internal representation independently given a rich enough label.

Varying model size exhibits predictable scaling. In Figure 2, we show results from a preliminary experiment regarding characterizing distillation in the physiological sensing domain. We first note that, consistent with prior work on the scaling laws for language model distillation [1], these trend lines follow a predictable exponential curve in mapping the size of student models to their distilled performance. We observe that performance seemingly begins to saturate at student models of size 6 residual blocks (139K parameters). We also note that although this figure shows fit curves for results obtained using the *soft distillation* strategy, the *DKD* and *feature distillation* strategies also adhered to these curves while the *hard distillation* strategy produced a much sharper saturation at an earlier point (i.e., at a smaller model size).

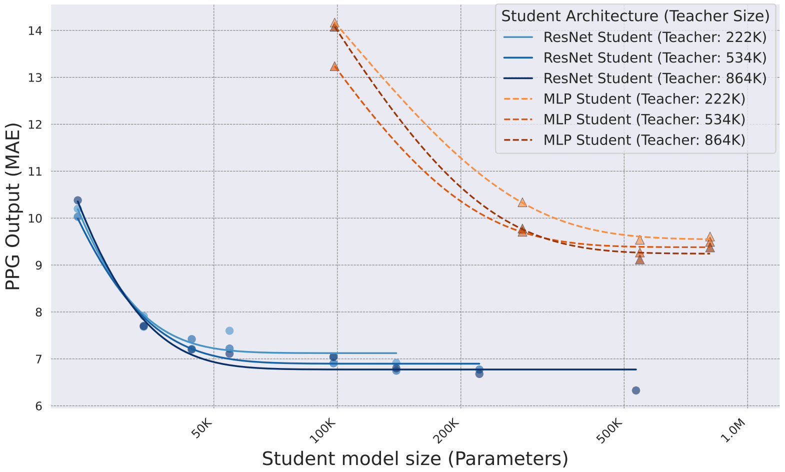

Figure 3: ***Soft* Distilled Model Scaling Behavior for ResNet and MLP Student Architectures.** Performance analysis of student models trained via soft distillation across varying parameter counts. ResNet students (blue) demonstrate superior scaling efficiency and a significantly lower error floor compared to MLP students (orange), indicating a stronger inductive bias for the PPG task.

Modeling decisions affect distillation scaling. We conduct an experiment to show the extent to which targeted modeling decisions affect distilled scaling by comparing the performance of distillation across different model architectures. Specifically, we compare the scaling behavior of distilled student models with ResNet and multilayer perceptron (MLP) backbones, while consistently distilling from a ResNet-like teacher model. Figure 3 illustrates how MLP-based models, similar to ResNet-based models, also exhibit predictable scaling as a function of both teacher and student capacities. However, we find that MLP-based student models consistently underperformed more sophisticated ResNet-based student models. We hypothesize that the innate inductive biases of the convolutional layers (e.g., a natural tendency to smoothly filter signals) paired with more targeted architecture designs such as residual connections enable more sample-efficient learning. We thus infer that while scaling may be observed for diverse model architecture, the specific scaling behavior may vary considerably.

Table 3: System compute benchmarks for distilled model inference. Inference time is reported in seconds (mean ± standard deviation) and peak GPU memory is reported in megabytes.

Time & Memory	**Model Size (Blocks)**								
Metrics	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**12**
Inference Time (s)	0.512 ±0.025	0.938 ±0.028	1.340 ±0.0316	1.787 ±0.144	2.177 ±0.192	2.622 ±0.167	3.357 ±0.147	4.419 ±0.115	4.758 ±0.130
Memory Usage (MB)	9.468	9.646	9.824	10.002	10.623	11.275	12.568	18.440	23.483

Distillation can lead to large gains in memory consumption and inference time. Table 3 shows system benchmarking of these models on an Nvidia RTX 2080-Ti GPU. Although this is not representative of our final application scenario (e.g., microprocessors in wearable devices), we include these results to show the relative improvement made possible by distillation. For example, distilling a 12-block model (i.e., the largest model we considered) to a 1-block model results in a nearly 90% decrease in inference time and 60% decrease in memory usage with only a 30% reduction in MAE performance.

4 Discussion and Conclusion

Dataset generalization. We presented an initial investigation into the distillation of heart rate estimation models. Our evaluation used a naive cross-validation scheme with shuffling samples from three datasets. We are interested in building off of work such as that of Kasnesis et al. [5] towards further studying the generalizability of these distilled models across datasets (i.e., by training on one dataset and testing on another).

Model architecture. Our preliminary investigation utilized a straightforward ResNet backbone model trained with supervision as the teacher model. We are interested in continuing our experiments using larger models trained with more recent contrastive approaches (e.g., we note that the model in [14] will be open source soon) to investigate how the potentially richer features learned in a self-supervised fashion might be distilled into smaller models.

Novel distillation strategies. In this work, we leveraged four approaches to distillation already documented in the literature to provide baseline characterizations of these heart rate estimation models. Leveraging insights from these experiments, we look forward to developing new methods of distillation that are particularly well-suited for this class of tasks.

This paper provides an initial demonstration of how knowledge distillation can be used to adapt large heart rate estimation models for resource-constrained wearable devices. Our preliminary evaluation shows that distilled models consistently outperform those trained from scratch, with *DKD* outperforming all other evaluated strategies. We also characterized a scaling law that confirms distillation enables substantial reductions in memory usage and inference time for a modest trade-off in performance. These findings provide an encouraging path forward for deploying powerful, real-time health monitoring models on the edge.

References

[1] Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

[2] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015.

[5] Panagiotis Kasnesis, Lazaros Toumanidis, Daniele Jahier Pagliari, and Alessio Burrello. Replacing attention with modality-wise convolution for energy-efficient ppg-based heart rate estimation using knowledge distillation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[6] Simon A Lee and Kai Akamatsu. Foundation models for physiological signals: Opportunities and challenges. 2025.

[7] Manuel Meier, Berken Utku Demirel, and Christian Holz. WildPPG: A real-world PPG dataset of long continuous recordings. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[8] 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.

[9] Sangjun Park, Dejiang Zheng, and Uichin Lee. A ppg signal dataset collected in semi-naturalistic settings using galaxy watch. *Scientific Data*, 12(1):892, 2025.

[10] Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel. Frill: A non-semantic speech embedding for mobile devices. *arXiv preprint arXiv:2011.04609*, 2020.

[11] Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024.

[12] Attila Reiss, Ina Indlekofer, and Philip Schmidt. PPG-DaLiA. UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C53890.

[13] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

[14] Mithun Saha, Maxwell A Xu, Wanting Mao, Sameer Neupane, James M Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *arXiv preprint arXiv:2502.01108*, 2025.

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[16] Muhammad Shabaan, Kaleem Arshid, Muhammad Yaqub, Feng Jinchao, M Sultan Zia, Giridhar Reddy Bojja, Muazzam Iftikhar, Usman Ghani, Loknath Sai Ambati, and Rizwan Munir. Survey: smartphone-based assessment of cardiovascular diseases using ecg and ppg analysis. *BMC medical informatics and decision making*, 20(1):177, 2020.

[17] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(1):1–30, 2021.

[18] Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

[19] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025.

[20] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation, 2022.