Sensor Applications

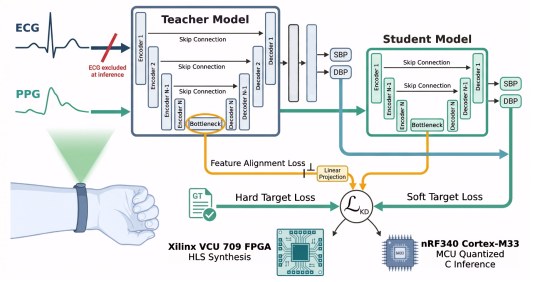
Compact PPG-Based Blood Pressure Estimation using Hardware-Efficient Knowledge Distillation

Neil Joshua Limbaga[1], Osama Amin[1], Basem Shihada[1] and Khaled Nabil Salama[1]

[1] *King Abdullah University of Science and Technology, Thuwal, 23955-3451, Kingdom of Saudi Arabia*

Manuscript received April 2026; revised July 12, 2026; date of current version July 12, 2026.

Abstract—Continuous blood pressure (BP) monitoring from photoplethysmography (PPG) is compelling for wearable health applications, yet deep learning models for cuffless BP estimation are rarely evaluated under edge-deployment constraints. We propose a hardware-oriented knowledge distillation framework for compact PPG-only BP estimation, where deployable students are trained using ground-truth BP labels, teacher predictions, and feature-level supervision from either PPG-only or hybrid ECG–PPG teachers. Evaluated on PhysioNet PTT and VitalDB, knowledge distillation via improves or maintains the accuracy–efficiency trade-off of compact students, especially in datasets with minimal data. A compact student model of a 1-depth 16-channel Residual UNet maintained a deployable size of 20.7k parameters with minimal penalty on average error. Hardware evaluation shows that the smallest, yet acceptable, student model fits within the VCU709 FPGA budget, while nRF5340 MCU deployment remains RAM-limited under full-window inference.



Index Terms—blood pressure estimation, photoplethysmography, knowledge distillation, edge deployment, wearable sensing

I. INTRODUCTION

Cardiovascular disease remains the leading cause of mortality worldwide, with hypertension affecting an estimated 1.4 billion adults globally with fewer than one in five having it adequately controlled [1]. Continuous blood pressure (BP) monitoring is a critical enabler of early intervention, yet conventional cuff-based sphygmomanometry is unsuitable for ambulatory use [2]. These limitations have driven research into cuffless BP estimation via photoplethysmography (PPG), which is low-cost and ubiquitous in wearables [3].

Despite rapid progress in deep learning for cuffless BP estimation [4], [5], deployment on resource-constrained hardware remains largely unaddressed. Recent hardware-oriented efforts synthesize an LSTM on a 90 nm ASIC (35.58 mm^2, 3.27 mW) using ECG and PPG inputs [6]; yet, custom silicon fabrication is inaccessible for most deployments, and ECG acquisition requires chest electrodes impractical for continuous wearable use. A recent survey identifies knowledge distillation (KD) as a principled compression strategy for edge-deployable BP monitoring [7], wherein a compact student model learns from both ground truth labels and the richer representations of a larger teacher model. KD has been demonstrated for hardware-constrained image inference on FPGA [8], yet its application to physiological signal-based BP estimation remains unexplored, particularly on micro-controller unit (MCU) platforms where on-chip SRAM and Flash are the binding deployment constraints [7].

In this paper, we propose a hardware-oriented knowledge distillation framework for compact PPG-only BP estimation, where deployable student models are trained using ground-truth BP labels, teacher predictions, and feature-level supervision from either PPG-only or hybrid ECG–PPG teachers. Unlike prior KD-based BP estimation models that retain ECG and PPG as inference inputs [9], and recent PPG-based KD work focused on heart-rate estimation rather than BP regression [10], the proposed framework targets PPG-only deployment and validates feasibility on FPGA and MCU platforms.

Corresponding author: N.J. Limbaga (e-mail: neil.limbaga@kaust.edu.sa).
Associate Editor: -
Digital Object Identifier 10.1109/LSENS.2023.0000000

II. METHODOLOGY

A. System Framework Overview

The framework consists of teacher training followed by student distillation, as shown in Fig. 1. Teacher models use either PPG-only or hybrid ECG–PPG inputs, while the deployed student is always a compact PPG-only UNet trained with ground-truth BP labels, teacher predictions, and feature-level supervision. ECG inputs, teacher networks, and feature-projection layers are used only during training and are removed from the deployed inference graph.

B. Multimodal Biosignal Datasets

We evaluate on two public datasets: (1) PhysioNet PTT [11], comprising synchronized 500-Hz PPG and ECG recordings from 22 subjects segmented into 30-s windows with activity-level SBP/DBP labels, and (2) VitalDB [12], comprising approximately 5,600 cases with continuous PPG, ECG, and arterial BP, resampled to 100 Hz and segmented into 10-s windows with 50% overlap. Subject-wise train/validation/test splits of 14/4/4 and 70/15/15 are applied before windowing for PTT and VitalDB, respectively, preventing subject leakage. All signals are z-score normalized per window.

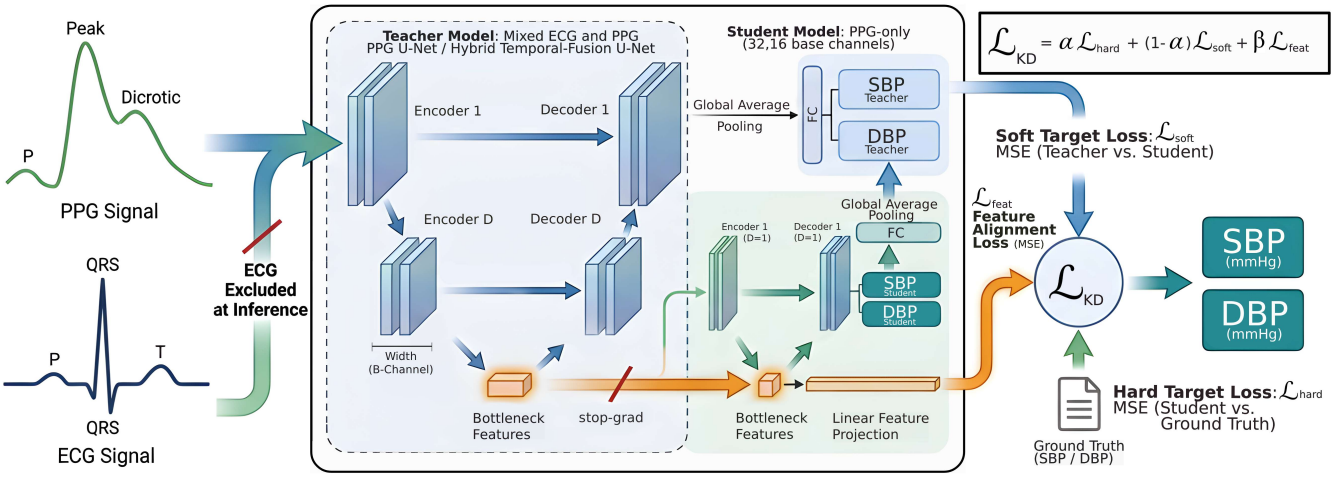


Fig. 1. Overview of the proposed knowledge distillation framework for compact PPG-based blood pressure estimation.

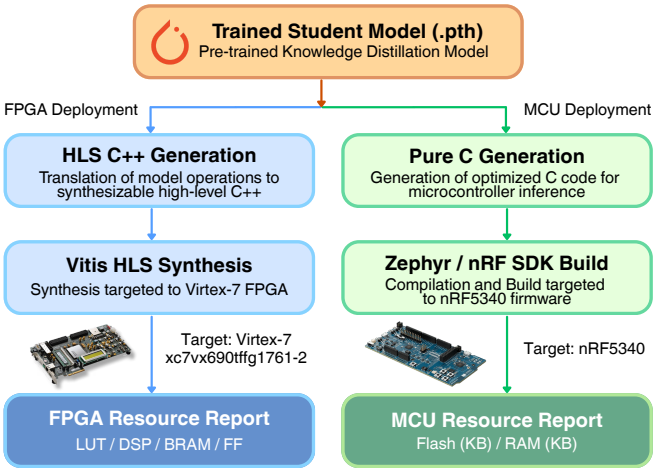


Fig. 2. Hardware deployment pipeline for distilled PPG-based blood pressure estimation. Starting from a trained PyTorch checkpoint, towards FPGA-deployment path (left) with target device Xilinx VCU709 (10 ns clock) and MCU-deployment path (right) with target device nRF5340 (Cortex-M33).

C. Knowledge Distillation Training

Both teacher and student are built on a 1D UNet backbone adapted for time-series regression [13]. Teacher architectures are selected through a controlled depth and width architecture search separately for two model structures: (1) the hybrid teacher, which contains modality-specific ECG and PPG encoders followed by temporal-fusion blocks; and (2) the PPG-only teacher, which uses a single PPG encoder. After teacher selection, compact PPG-only students are trained with candidate widths and optimization settings. Feature-level distillation is applied at the bottleneck, where the temporal representation is most compact after encoder downsampling. When the student and teacher bottleneck dimensions differ, a lightweight linear projection maps the student bottleneck from dimension B_s to the teacher bottleneck dimension B_t before computing the feature-alignment loss. The teacher is frozen and the student optimized with:

$$\mathcal{L}_{\mathrm{KD}} = \alpha \mathcal{L}_{\mathrm{hard}} + (1-\alpha)\mathcal{L}_{\mathrm{soft}} + \beta \mathcal{L}_{\mathrm{feat}}, \quad (1)$$

where $\mathcal{L}_{\mathrm{hard}}$ is MSE against ground-truth BP, $\mathcal{L}_{\mathrm{soft}}$ is MSE between student and teacher predictions, and $\mathcal{L}_{\mathrm{feat}}$ aligns the projected bottleneck features with a stop-gradient on the teacher branch [14]. The coefficient β controls the contribution of feature-level distillation and is selected on the validation set. In the student objective, α is set to 0.5, allowing equal importance for $\mathcal{L}_{\mathrm{hard}}$ and $\mathcal{L}_{\mathrm{soft}}$. Models are trained with Adam (learning rate = 10^{-3}, weight decay 10^{-5}) with early stopping on the validation set's Mean Absolute Error (MAE).

D. Hardware Deployment

To evaluate deployability across edge hardware targets, we assess each model on both an FPGA and MCU as illustrated in Fig. 2.

FPGA Synthesis. For FPGA evaluation, each trained model is translated into HLS C++ and synthesized for the Xilinx VCU709 with 16-bit fixed-point weights and a 10 ns clock constraint. The generated design preserves the 1D UNet inference graph, including convolution, pooling, upsampling, skip connections, and regression layers, and reports BRAM, DSP, FF, LUT, and latency estimates.

MCU Deployment. For MCU evaluation, we target the Nordic nRF5340 with Zephyr RTOS. Batch-normalization parameters are folded into convolution weights, model weights are symmetrically quantized to 8-bit integers, and a dependency-free C inference engine with static activation buffers is used to estimate Flash and RAM requirements.

III. EXPERIMENTS AND RESULTS

This section shows the evaluation across two areas: (1) BP estimation performance and efficiency using knowledge distillation, and (2) FPGA and MCU resource utilization.

A. Distillation Results

For the architecture sweep as shown in Fig. 3, both the PPG-only Residual U-Net and the hybrid ECG–PPG temporal-fusion model were evaluated across encoder–decoder depths and base-channel widths. The sweep shows that shallow PPG-only models generally provided the strongest and most stable performance, with $D = 1$ selected for the PPG-only teacher on both datasets. For the hybrid ECG–PPG teacher, the selected depth was dataset-dependent: $D = 1$

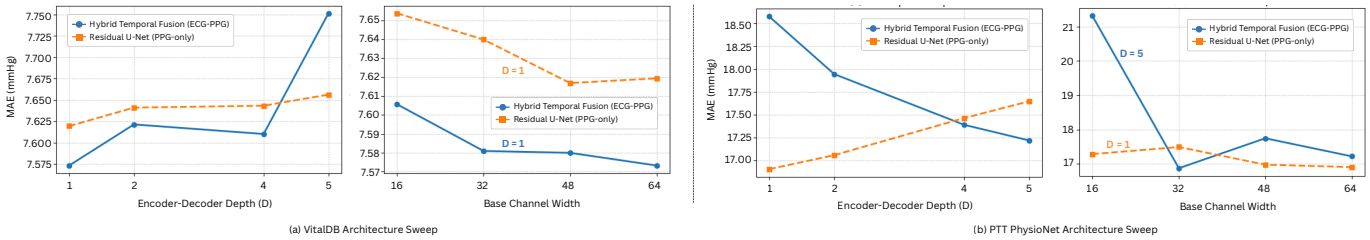


Fig. 3. Architecture sweep for the PPG-only and Hybrid Teacher Models on VitalDB (a) and (b) PhysioNet PTT.

TABLE 1. Blood pressure estimation results on the PTT and VitalDB datasets. MAE values are reported in mmHg; Size in MB using float32.

Model	Input	Params (K)	Size (MB)	PTT (PhysioNet) SBP	PTT DBP	PTT Avg.	PTT FOM	VitalDB SBP	VitalDB DBP	VitalDB Avg.	VitalDB FOM
Prior Work											
MHSA-AE [5]	PPG+VPG+APG	79348*	317.39*	–	–	–	–	7.540	**5.130**	6.335	31.543
XResNet1d101 [15]	PPG	28273*	113.09*	–	–	–	–	9.400	6.000	7.700	22.635
Direct PPG-only Models											
Direct PPG-UNet (D1/32)	PPG	81.794	0.314	22.195	12.790	17.492	6.653	9.064	6.215	7.640	34.874
Direct PPG-UNet (D1/16)	PPG	20.674	0.082	22.483	12.090	17.287	7.755	9.089	6.219	7.654	39.557
Distilled PPG-only Students (PPG Teacher)											
Student PPG-UNet, $\beta = 0.10$ (D1/32)	PPG	81.794	0.314	**21.880**	**12.200**	**17.040**	7.010	**9.048**	**6.184**	**7.616**	35.092
Student PPG-UNet, $\beta = 0.50$ (D1/16)	PPG	20.674	0.082	23.135	12.769	17.952	7.190	9.085	6.209	7.647	**39.624**
Distilled PPG-only Students (Hybrid ECG–PPG Teacher)											
Student PPG-UNet, $\beta = 0.10$ (D1/32)	PPG	81.794	0.314	22.886	12.850	17.868	6.376	**9.055**	**6.185**	**7.620**	35.057
Student PPG-UNet, $\beta = 0.50$ (D1/16)	PPG	20.674	0.082	**22.052**	**12.332**	**17.192**	**7.840**	9.097	6.215	7.656	39.533

*Parameter counts reconstructed from published architecture descriptions. **Bold**: best value within each distilled student group. Underline: best FOM per dataset.

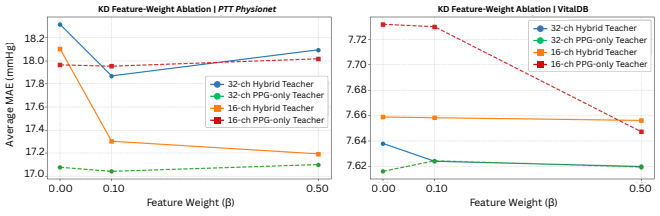


Fig. 4. Feature-distillation weight ablation for the selected PPG-only and hybrid ECG–PPG teachers on VitalDB and PhysioNet PTT.

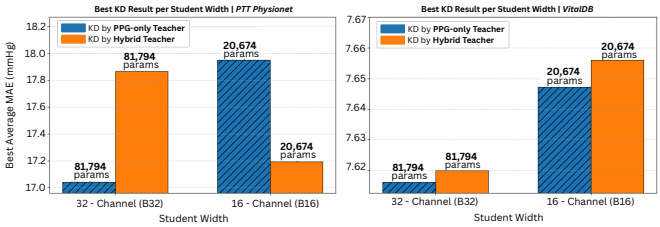


Fig. 5. Accuracy–efficiency trade-off of the selected direct and distilled PPG-only deployment models on the PTT and VitalDB datasets.

on VitalDB and $D = 5$ on PhysioNet PTT. These selected teacher architectures were then used for the subsequent KD experiments, while the deployed student remained PPG-only.

After selecting the teacher architectures from the depth–width sweep, we further ablate the feature weight coefficient β in the KD objective. Fig. 4 shows the average MAE for the compact student models with a depth of 1 and base channels of 32 and 16 (D1/32 and D1/16) obtained using $\beta \in \{0, 0.1, 0.5\}$ via PPG-only and hybrid ECG–PPG teacher models. In most settings, including a nonzero feature-alignment term improves or maintains the average MAE relative to output-only distillation. This is particularly evident on the limited PhysioNet PTT dataset, where the hybrid-teacher D1/16 student improves substantially when β is nonzero, indicating that intermediate feature supervision can provide useful structural guidance for compact PPG-only deployment models.

To compare accuracy–compactness trade-offs, we define a parameter-aware figure of merit as

$$\mathrm{FOM} = \frac{10^4}{\bar{E}^2 \log_{10}(N_{\mathrm{param}})}, \quad \bar{E} = \frac{\mathrm{MAE}_{\mathrm{SBP}} + \mathrm{MAE}_{\mathrm{DBP}}}{2}. \tag{2}$$

where N_{param} denotes the number of deployable parameters. Higher FOM values indicate a better accuracy–compactness trade-off. The squared error term emphasizes small changes in MAE, while the logarithmic parameter term accounts for the fact that model sizes vary across orders of magnitude and prevents parameter count from dominating the metric linearly. The factor 10^4 is used only to improve numerical readability and does not affect model ranking.

Table 1 summarizes the final accuracy–efficiency trade-off of the selected direct and distilled PPG-only deployment models on the PTT and VitalDB datasets. On VitalDB, the direct shallow PPG-only models already provide strong performance, so KD yields only marginal gains, with the best distilled results remaining very close to the direct baselines. In contrast, the PTT dataset shows a clearer benefit from teacher-guided compression. In particular, distillation from the hybrid ECG–PPG teacher enables a compact D1/16 student with only 20.7k deployment parameters to achieve an average MAE of 17.1918 mmHg, improving over the direct D1/16 PPG baseline (17.2865 mmHg) while also increasing the FOM from 7.7547 to 7.8403. As seen in Fig. 5, the 16-channel student distilled from the hybrid teacher may have a slightly higher MAE than the 32-channel student distilled from the PPG-only teacher; but has a reduction of 73.5%, marking the best FOM for a student model. This indicates that multimodal teacher supervision is most

TABLE 2. Hardware resource reports for the deployment models across FPGA and MCU deployment targets.

					FPGA (VCU709)										MCU (nRF5340)			
					BRAM		DSP		FF		LUT		Latency		Flash		RAM	
Model	Width	Params.	Avg. MAE	FOM	Count	%	Count	%	Count	%	Count	%	M cycles	ms	KB	%	KB	%
Direct PPG	D1/32	81,794	17.4923	6.65	5614	190.95	137	3.81	10331	1.19	13789	3.18	1054.80	10548.00	81.12	7.92	12187.75	2380.42
PPG-KD, $\beta = 0.10$	D1/32	81,794	**17.0401**	**7.01**	5614	190.95	137	3.81	10328	1.19	13785	3.18	1054.80	10548.00	81.12	7.92	12187.75	2380.42
Direct PPG	D1/16	20,674	17.2865	7.75	2797	95.14	105	2.92	9000	1.04	12878	2.97	266.28	2662.80	20.84	2.03	6093.88	1190.21
Hybrid-KD, $\beta = 0.50$	D1/16	20,674	**17.1918**	**7.84**	2798	95.17	105	2.92	8967	1.03	12880	2.97	266.28	2662.80	20.84	2.03	6093.88	1190.21

Avg. MAE (PTT Results). Latency is reported in million cycles and milliseconds at 100 MHz. For the nRF5340 target, flash utilization is computed relative to 1 MB flash and RAM utilization is computed relative to 512 KB RAM. Green denotes resources within the target budget, while red denotes resources exceeding the target budget.

beneficial in the more challenging low-parameter regime, where it helps preserve predictive performance while maintaining a compact unimodal deployment model. For prior methods, parameter counts were reconstructed from the architectural descriptions reported in the corresponding papers, including the number of layers, channel widths, kernel sizes, and fully connected layers when available. These reconstructed counts are therefore approximate and are marked with an asterisk in Table 1. Because implementation details such as padding conventions, normalization layers, or auxiliary modules may not be fully specified, comparisons involving prior-work parameter counts should be interpreted as approximate complexity estimates only.

B. Edge Deployment Feasibility

As seen from Table 2, the hardware benefit of KD is not obtained by changing the deployed architecture, but by improving the accuracy–efficiency trade-off at the same PPG-only student width. For a fixed width, the direct and distilled students have nearly identical FPGA utilization because the teacher network and feature-projection layer are used only during training and are removed from the deployed graph. The D1/32 models provide stronger accuracy but exceed the BRAM budget of the VCU709 target, reaching approximately 191% BRAM utilization. In contrast, the D1/16 models fit within the FPGA budget, requiring about 95% BRAM utilization while keeping DSP, FF, and LUT usage below 4%. Notably as mentioned, the student model at D1/16 distilled by the hybrid-teacher model has the best FOM, which allowed a 74.76% decrease in latency.

For MCU deployment, the limiting factor is RAM rather than flash. All selected models fit comfortably within the 1 MB flash budget of the nRF5340, requiring only 81.12 KB for D1/32 and 20.84 KB for D1/16 after weight quantization. However, the direct full-window implementation exceeds the 512 KB RAM budget because the activation buffers scale with the 15,000-sample input window. The D1/16 models require 6093.88 KB RAM, while the D1/32 models require 12187.75 KB RAM. Thus, the selected D1/16 hybrid-KD model is feasible on the FPGA target and provides the best compact-model trade-off, but MCU deployment would require additional memory optimization such as buffer reuse or streaming inference.

IV. CONCLUSION

This paper presented a hardware-oriented knowledge distillation framework for compact PPG-based blood pressure estimation. The proposed framework trains deployable PPG-based student models using ground-truth BP labels, teacher predictions, and feature-level supervision from either PPG-only or hybrid ECG–PPG teachers. Across PhysioNet PTT and VitalDB, the results show that KD improves or maintains the accuracy–efficiency trade-off of compact students, with the strongest benefit observed in the low-parameter PTT setting. In particular, the hybrid-teacher D1/16 student achieved the best compact-model trade-off on PTT, improving average MAE and FOM over the direct PPG baseline without increasing deployed model size. Hardware evaluation further showed that the D1/16 student fits within the VCU709 FPGA resource budget, whereas MCU deployment on the nRF5340 remains RAM-limited under full-window inference. Future work will explore more memory-efficient inference schedules and higher-capacity teacher architectures for improved cross-modal supervision.

REFERENCES

[1] World Health Organization, "Global report on hypertension 2025: uncontrolled high blood pressure puts over a billion people at risk," WHO, Geneva, Switzerland, Sep. 2025. [Online].

[2] R. Mukkamala, S. G. Shroff, K. G. Kyriakoulis, A. P. Avolio, and G. S. Stergiou, "Cuffless blood pressure measurement: where do we actually stand?" *Hypertension*, vol. 82, no. 6, 2025, doi: 10.1161/HYPERTENSIONAHA.125.24822.

[3] C. M. Y. Cheung, A. Sabharwal, G. L. Coté, and A. Veeraraghavan, "Wearable blood pressure monitoring devices: understanding heterogeneity in design and evaluation," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 12, pp. 3569–3592, Dec. 2024.

[4] T. Panula, J. P. Sirkia, D. Wong, and M. Kaisti, "Advances in non-invasive blood pressure measurement techniques," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 424–438, 2023.

[5] H. Rong, H. Wu, and R. Liu, "An autoencoder-based semi-supervised learning for efficient blood pressure estimation with minimal labeled data," in *Proc. IEEE AICAS*, 2025.

[6] T. Joseph and T. S. Bindiya, "Real-time blood pressure prediction on wearables with edge-based DNNs: a co-design approach," *ACM Trans. Design Autom. Electron. Syst.*, vol. 30, no. 1, art. no. 11, Dec. 2024.

[7] Y. Zhang, S. Qiu, K. Du, S. Wu, T. Xiang, K. Zheng, Z. Liu, H. Chen, N. Ji, F. Wang, W. Wu, and Y.-T. Zhang, "Artificial intelligence-enhanced wearable blood pressure monitoring in resource-limited settings: a co-design of sensors, model, and deployment," *Nano-Micro Lett.*, vol. 18, art. no. 164, 2026.

[8] L. Zhang, A. M. Eltawil, and K. N. Salama, "Hardware-friendly lightweight convolutional neural network derivation at the edge," in *Proc. IEEE 6th Int. Conf. AI Circuits Syst. (AICAS)*, Abu Dhabi, UAE, Apr. 2024, pp. 11–15. DOI: 10.1109/AICAS59952.2024.10595961.

[9] H. Tang, G. Ma, L. Qiu, L. Zheng, R. Bao, J. Liu, and L. Wang, "Blood pressure estimation based on PPG and ECG signals using knowledge distillation," *Cardiovasc. Eng. Technol.*, vol. 15, pp. 39–51, 2024, doi: 10.1007/s13239-023-00695-x.

[10] K. Arora, G. Narayanswamy, S. Patel, and R. Li, "Towards characterizing knowledge distillation of PPG heart rate estimation models," in *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025.

[11] P. Mehrgardt, M. Khushi, S. Poon, and A. Withana, "Pulse Transit Time PPG Dataset (version 1.1.0)," *PhysioNet*, 2022. https://doi.org/10.13026/jpan-6n92

[12] H.-C. Lee, Y. Park, S. B. Yoon et al., "VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients," *Scientific Data*, vol. 9, no. 279, 2022. https://doi.org/10.1038/s41597-022-01411-5

[13] S. Mahmud et al., "A shallow U-Net architecture for reliably predicting blood pressure (BP) from photoplethysmogram (PPG) and electrocardiogram (ECG) signals," *Sensors*, vol. 22, no. 3, p. 919, 2022.

[14] D. Qin et al., "Efficient medical image segmentation based on knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3820–3831, 2021.

[15] M. Moulaeifard, P. H. Charlton, and N. Strodthoff, "Generalizable deep learning for photoplethysmography-based blood pressure estimation: A benchmarking study," *Mach. Learn.: Health*, vol. 1, p. 010501, 2025.