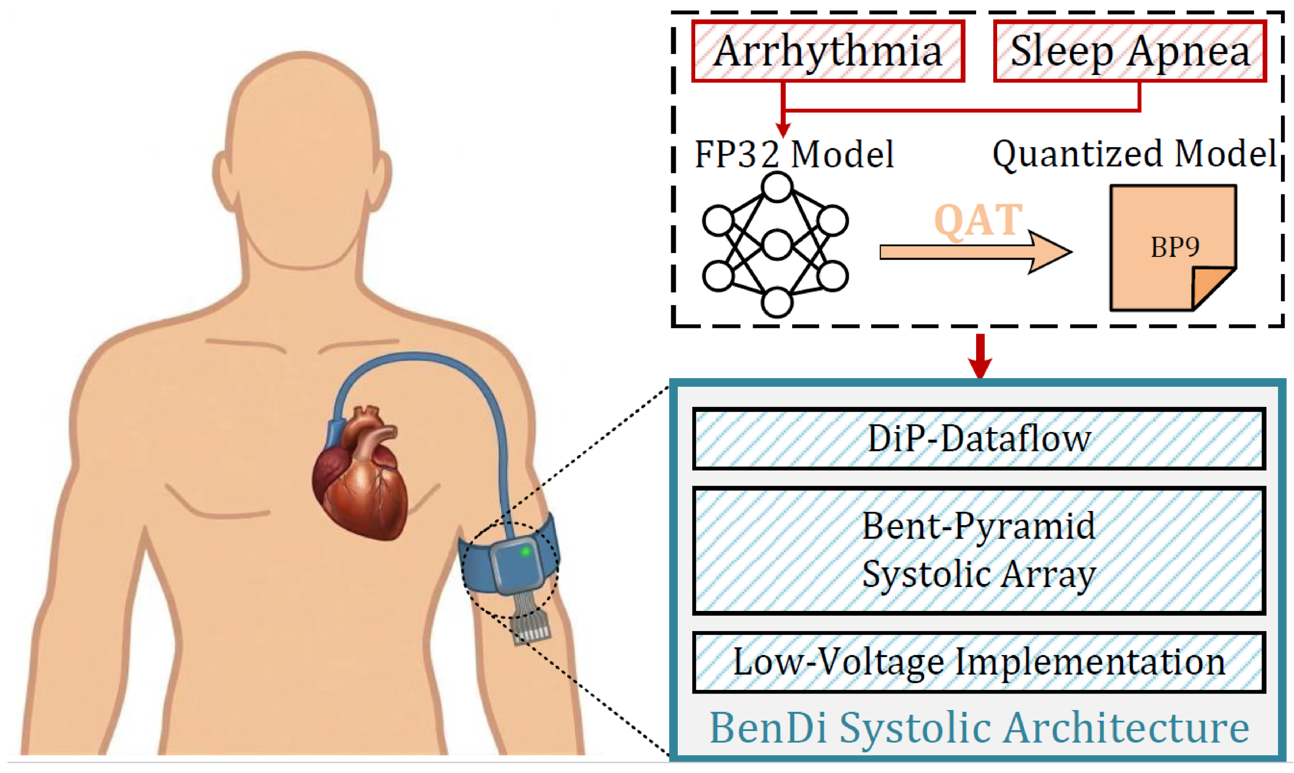
BenDi: An Energy-Efficient Quasi-Stochastic Systolic Architecture for Edge Bioelectronics

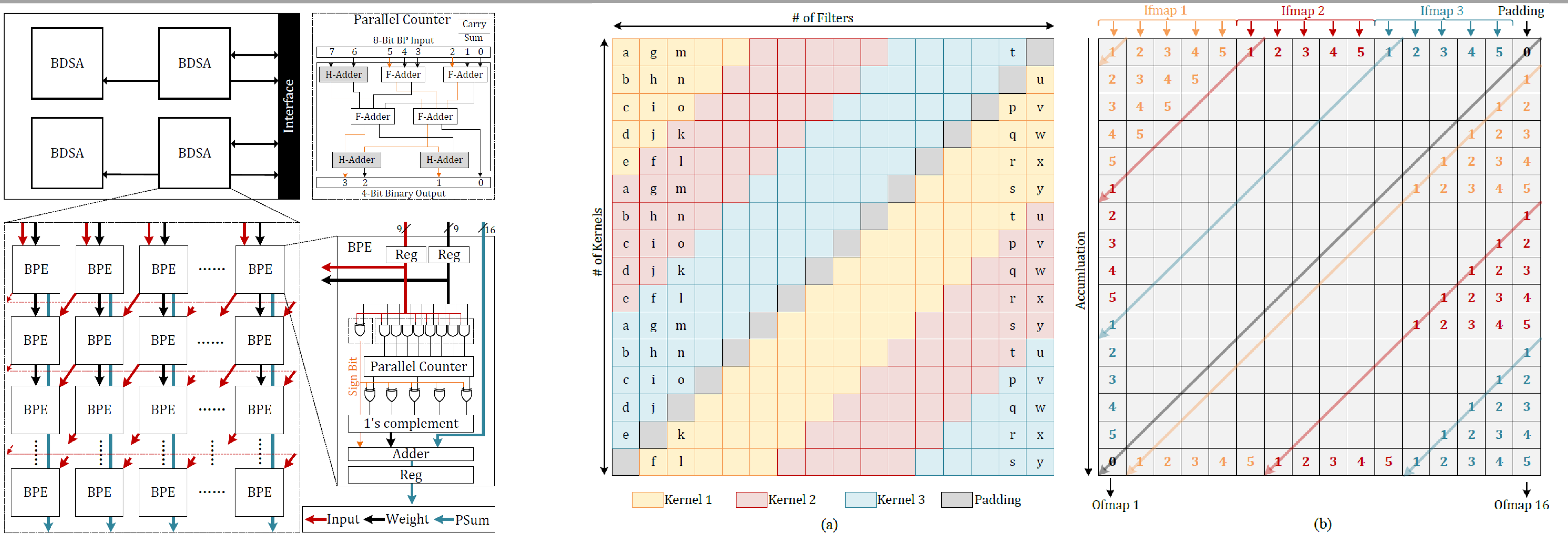
Bochen Ye†, Yihan Pan†, Shady Agwa†, Themis Prodromakis†
†Centre for Electronics Frontiers (CEF), Institute for Integrated Micro and Nano Systems (IMNS)

Introduction

Continuous long-term monitoring and diagnosis of biomedical signals, such as electrocardiograms (ECGs), can help mitigate an increasing threat to public health. Artificial Intelligence (AI) models, such as Convolutional Neural Networks (CNNs), provide accurate monitoring and classification for relevant diseases; however, they require more computational resources than conventional AI hardware can typically afford, especially for a resource-constrained environment on the edge.



BenDi Architecture



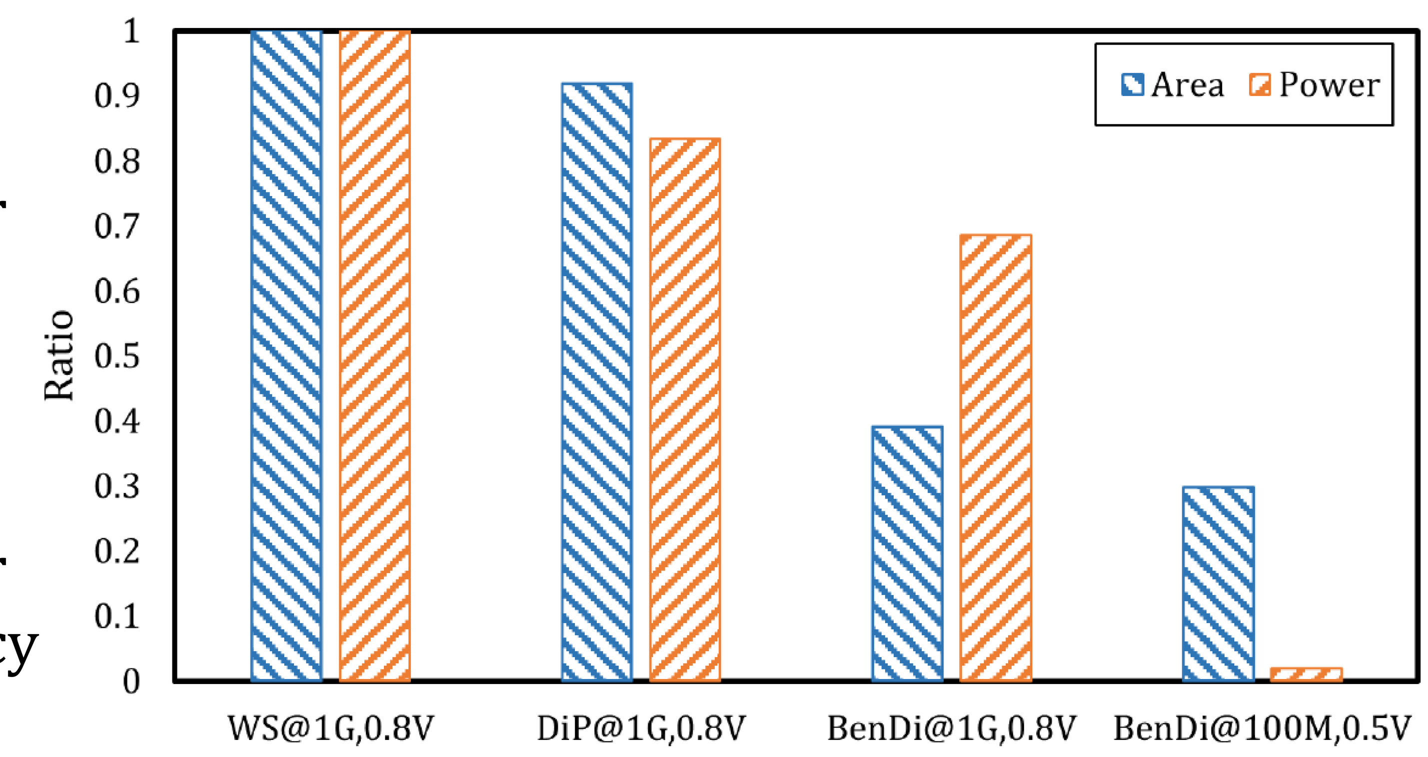
Hardware Results

- Per-layer utilization/latency/energy

	MIT-BIH			Apnea-ECG		
	Util.	Latency	Energy	Util.	Latency	Energy
Conv1	25%	1.28	1.64	50%	4.8	12.31
Conv2	100%	3.36	17.23	100%	9.6	49.24
FC1	100%	4	20.51	100%	7.84	40.21
FC2	50%	0.48	2.46	50%	0.48	1.23
Total	**86.84%**	**9.12**	**41.84**	**88.38%**	**22.72**	**102.99**

Note: Latency is measured in μs. Energy is measured in nJ.

- @ 1 GHz, 0.8 V
  2.56x smaller area
  1.45x lower power
  Vs. WS
- @ 100 MHz, 0.5 V
  3.35x smaller area
  50.1x lower power
  5x energy-efficiency
  Vs. WS



- Compare with prior work

	[1]	[2]	**BenDi**
Paradigm	Binary	Binary	**Bent-Pyramid**
Model	1D-CNN	BNN	1D-CNN
Dataset	MIT-BIH	MIT-BIH	MIT-BIH/Apnea-ECG
Data Width	16-bit	1-bit	**9-bit**
Accuracy(%)	98.4	98.78	**96.89**/84.31
Frequency (Hz)	100M	100K	**100M**
Process (nm)	55	55	**22**
Voltage (V)	0.85	1.2	**0.5**
Area (mm^2)	1.39 (0.305*)	0.43 (0.204*)	**0.085**
Power (mW)	5.14 (3.212*)	0.00784 (0.00482*)	**5.1299**
Throughput (GOPS)	2.04	-	**204.8**
Area Eff. (TOPS/mm^2)	0.001 (0.099*)	-	**2.409**
Energy Eff. (TOPS/W)	0.397 (0.635*)	-	**39.92**

*: Technology scaling to 22nm according to DeepScaleTool [13].

Software Results

- Accuracy Result: 1%-3.3% loss

Methods	MNIST	MIT-BIH		Apnea-ECG	
	ACC	Macro-F1	ACC	Macro-F1	ACC
FP32	98.84%	0.90	97.89%	0.871	87.59%
PTQ-INT8	98.63%	0.824	96.57%	0.754	77.28%
PTQ-BP9	97.45%	0.701	91.66%	0.771	77.14%
QAT-BP9	-	**0.854**	**96.89%**	**0.829**	**84.31%**

[1] C. Zhang, Z. Huang, C. Zhou, A. Qie, and X. Wang, "An energy-efficient configurable 1-d cnn-based multi-lead ecg classification coprocessor for wearable cardiac monitoring devices," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 19, no. 2, pp. 317–331, 2025.
[2] R. Zhang, R. Zhou, X. Han, H. Qi, and Y. Wang, "An energy-efficient ecg classifier with on-chip learning using binarized convolutional neural network," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 20, no. 1, pp. 107–121, 2026.
[3] A. J. Abdelmaksoud, S. Agwa, and T. Prodromakis, "Dip: A scalable, energy-efficient systolic array for matrix multiplication acceleration," IEEE Transactions on Circuits and Systems I: Regular Papers, pp. 1–11, 2025.

Conclusion

1. BenDi improves area and energy efficiency through multi-level optimization: hardware-aware BP9 quantization preserves accuracy, DiP dataflow[3] reduces synchronization-buffer overhead, quasi-stochastic computing simplifies multipliers, and low-voltage implementation further reduces power.
2. BenDi achieves 3.35× smaller area and 5× higher energy efficiency than a conventional WS systolic array, with only 1%–3.3% accuracy loss versus FP32.
3. Compared with the prior bioelectronics accelerator, BenDi achieves 24.33× higher area efficiency and 62.86× higher energy efficiency.