Hardware-Accelerated Deep Learning for Multimodal Biomedical Monitoring With On-Device Adaptive Learning on Resource-Constrained SoC

Sara Aly

Department of Computer Engineering, California State University, Northridge

Advisor: Dr. Shahnam Mirzaei | Accepted at IEEE ICHI 2026

1 MOTIVATION / QUESTION OF INTEREST

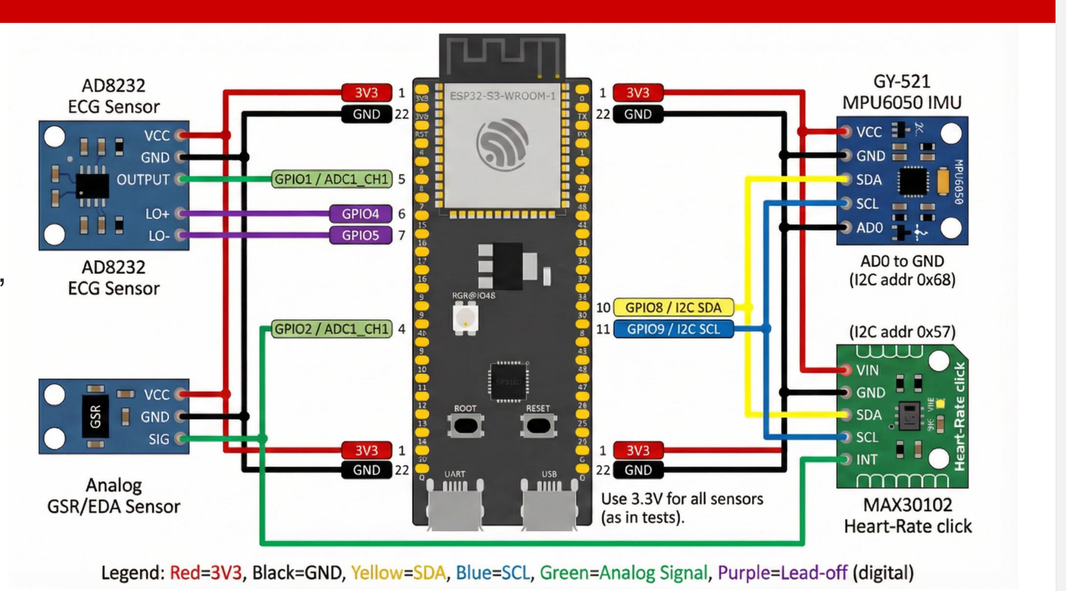
Can a sub-$15 microcontroller simultaneously monitor physical activity, psychological stress, and cardiac arrhythmia from body-worn sensors — and adapt its predictions to individual users over time — entirely on-device, without cloud connectivity or external ML frameworks?

- Cardiovascular disease is the #1 cause of death globally; arrhythmias like atrial fibrillation are often asymptomatic and missed during brief clinic visits — continuous ambulatory monitoring is the only way to catch them
- Chronic stress contributes to hypertension and immune dysfunction, but real-time stress detection outside the lab remains unsolved
- Existing solutions: Apple Watch ($400+) single-lead ECG, single-task only; clinical Holter monitors ($500+) require clinic visits; cloud-dependent systems have latency and privacy issues. Our system: ~$45 total hardware cost
- A single wearable monitoring stress, arrhythmia, AND activity simultaneously can reveal cross-domain interactions (stress-induced arrhythmia, exercise-modulated anxiety) invisible to single-task systems

2 BACKGROUND

Hardware Platform

- ESP32-S3-DevKitC-1 (N8R8): Xtensa LX7 dual-core 240 MHz, 520 KB SRAM, 8 MB PSRAM, 8 MB Flash — retail ~$45
- Sensors: AD8232 (single-lead ECG), MAX30102 (PPG/SpO2), MPU6050 (3-axis IMU), optional GSR (analog)



Datasets

- PPG-DaLiA (Reiss et al., 2019): 15 subjects, wrist PPG + chest ECG/ACC, activity labels → 12,806 windows
- WESAD (Schmidt et al., 2018): 15 subjects, chest + wrist sensors, 3-class stress → 3,439 windows
- MIT-BIH Arrhythmia (Moody & Mark, 2001): 47 subjects, 2-lead ambulatory ECG → 5,459 windows
- Unified corpus: 21,704 windows from 77 subjects; 5-channel representation (ECG, PPG, AccX, AccY, AccZ)

3 OBJECTIVE

Design, train, deploy, and validate an end-to-end multimodal biomedical monitoring system:

- Joint classification of activity (4 classes), stress (binary), and arrhythmia (binary) from a single 10-second window
- Full model deployed on ESP32-S3 without quantization or pruning — training model = deployed model
- On-device adaptive training for per-user personalization with 9 safety mechanisms
- Bare-metal C++ (~1,200 lines), zero dynamic allocation, no RTOS, no external ML libraries

Four Context-Aware Clinical Alert Scenarios

Case	Detected Condition	Activity State	Alert Action	Rationale
1	Stress	Sedentary	psychological_stress	Genuine stress at rest
2	Stress	Exercise	no_alert (suppressed)	Elevated stress during exercise is expected
3	Arrhythmia	Sedentary	arrhythmia_detected	Highest clinical priority
4	Arrhythmia	High Motion	critical_alert	Potentially life-threatening

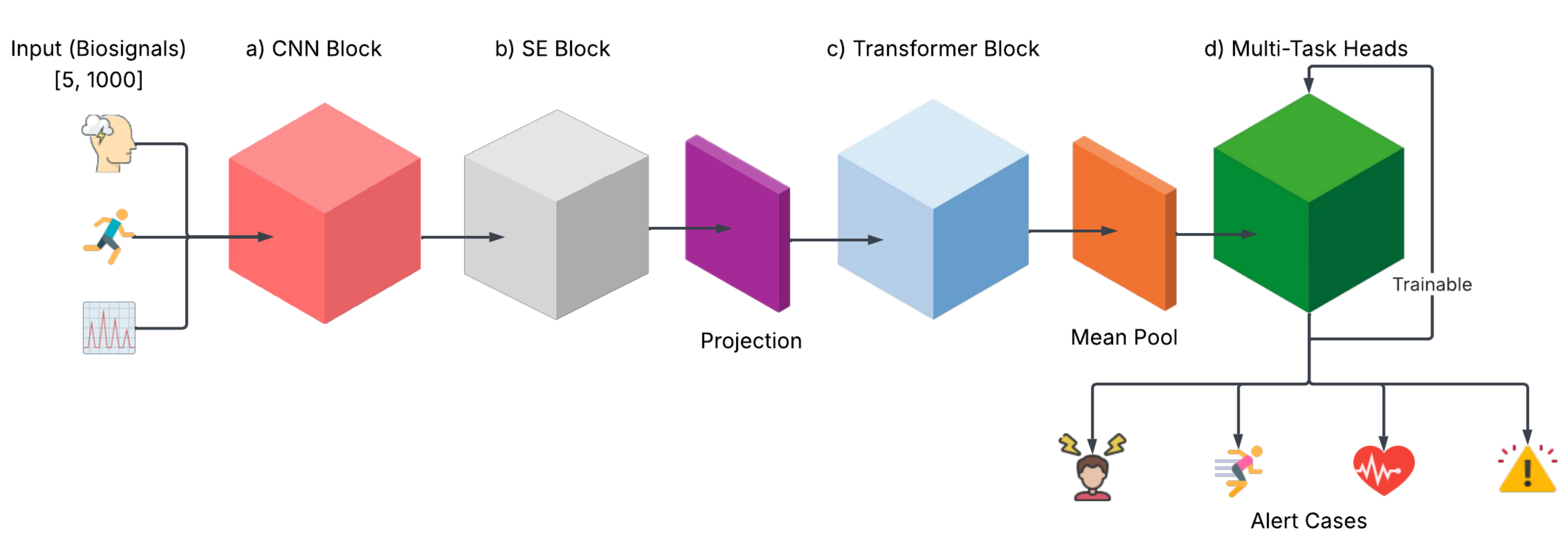


Figure 1. Total: 112,552 parameters (450 KB FP32) — deployed uncompressed on ESP32-S3

4 METHODOLOGY

Model Architecture: CNN-SE-Transformer (V5)

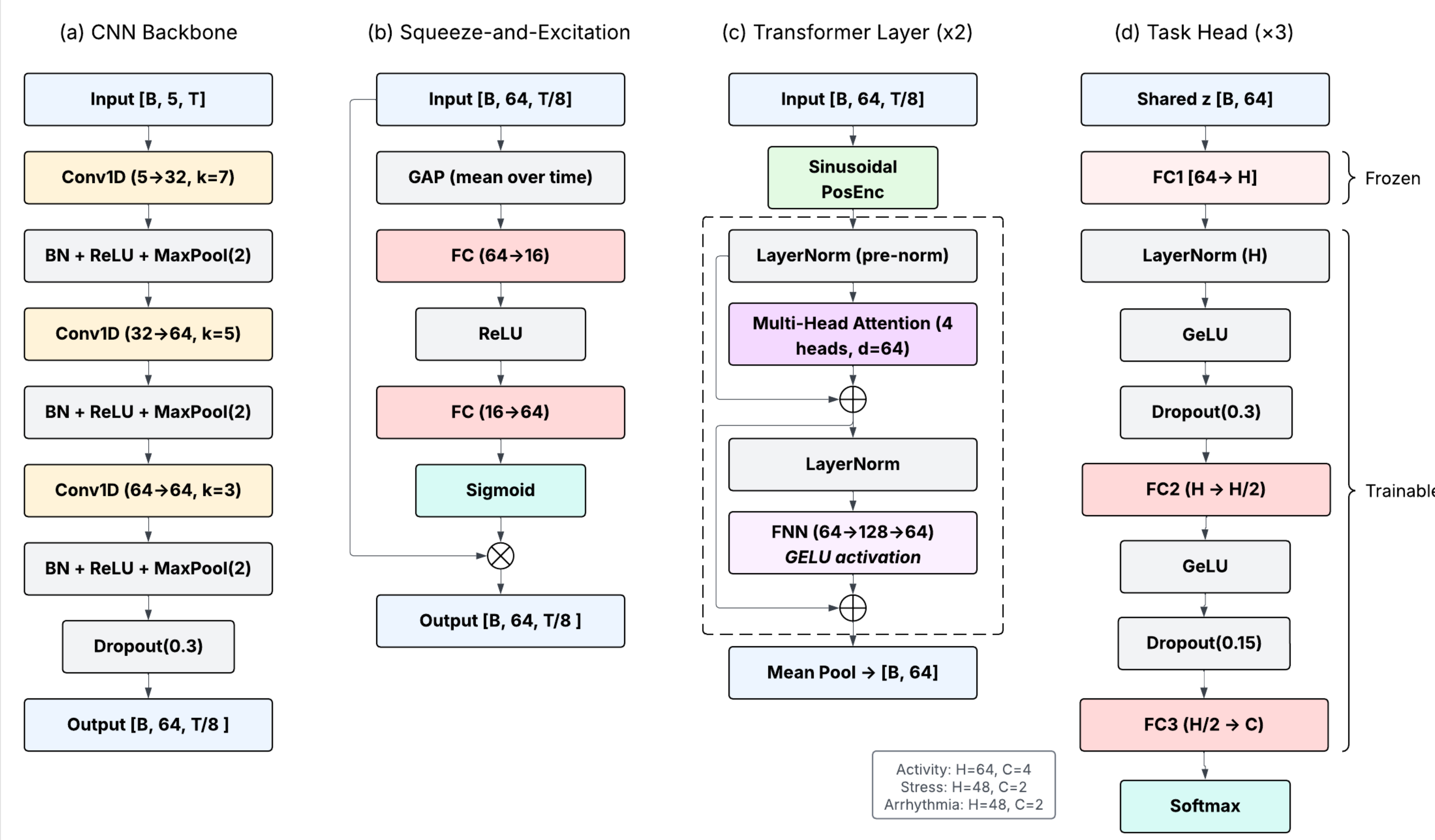


Figure 2. Component architectures: (a) CNN backbone with BN fusion, (b) SE attention, (c) Transformer layer, (d) 3-layer classification heads

Training Strategy

- Subject-wise splits: 46 train / 6 validation / 13 test (10,134 windows — zero overlap with training data)
- Class-balanced focal loss with per-task weighting + weighted random sampling to counter severe class imbalance
- Three-phase fine-tuning: (1) all heads joint, (2) arrhythmia head alone, (3) stress head alone
- Augmentation: random scaling (±10%), temporal shifting, Gaussian noise injection

5 ON-DEVICE ADAPTIVE TRAINING

Architecture

- Frozen backbone: 107,568 params (CNN + SE + Transformer + head FC1 layers) — never updated on device
- Trainable: 4,984 params (LayerNorm + FC2 + FC3 per head) — head-only backprop reduces activation storage 715× vs. full backprop
- Forward: FC1(frozen) → LN → GELU → FC2 → GELU → FC3 → Softmax → Cross-Entropy loss
- Backward: gradients through FC3, FC2, LayerNorm only; SGD with momentum (lr=0.05, μ=0.9), gradient clipping at 1.0

Nine Safety Mechanisms

- (1) Gradient clipping (2) Weight clamping (3) NaN/Inf guards (4) A/B validation gate: candidate promoted only if accuracy ≥ stable model on all 3 tasks
- (5) Weight divergence scan (6) Safety lockout after consecutive failures (7) Thermal guard (T < 65°C) (8) Wall-clock budget (< 500 ms)
- (9) Factory reset to original weights via NVS flash — dual-slot model versioning persists across reboots
- Resource gating: requires PSRAM ≥ 2 MB free, heap ≥ 50 KB free, inference latency ≤ 3 sec before allowing training

On-Device Training Comparison

System	Platform	Tasks	Training Time	Safety Mech.	Personalization
TinyOL	Arduino Nano 33 BLE	1	~100 ms	0	Output layer
TinyTL	Cortex-A	1	~1 sec	0	Bias-only
TinyTrain	Cortex-A	1	~2 sec	1	Sparse update
This Work	**ESP32-S3**	**3 (joint)**	**16 ms**	**9**	**Head-only + A/B**

16 ms Mean Training Episode | **+8 KB** Flash Overhead (+0.2%) | **4,984** Trainable Params (of 112,552) | **715×** Activation Storage Reduction

6 RESULTS & ANALYSIS

Multi-Task Classification on Held-Out Test Set

Task	Acc. [95% CI]	AUC [95% CI]	F1-macro	Sens.	Spec.
Activity (4-class)	94.1% [93.5, 94.8]	0.731	0.259	—	—
Stress (binary)	87.1% [85.3, 88.8]	0.975 [0.965, 0.984]	0.819	66.2%	97.5%
Arrhythmia (binary)	90.8% [89.9, 91.8]	0.879 [0.862, 0.894]	0.646	36.0%	94.7%

Arrhythmia: high specificity (94.7%) = low false alarm rate, suitable for clinical screening. Stress AUC 0.975 = near-perfect discrimination.

Real Hardware Measurements (ESP32-S3-N8R8, all sensors active)

2,338 ms Inference Latency per 10-sec Window | **0.54 W** Power (5.23V × 0.10A) | **1,263 mJ** Energy per Inference | **23.4%** Duty Cycle (idle 76.6%)

Flash: 768 KB (23.0%) | SRAM: 86 KB (26.2%)
Inference buffers: 687 KB in PSRAM (7.3 MB free) | Heap min: 288 KB

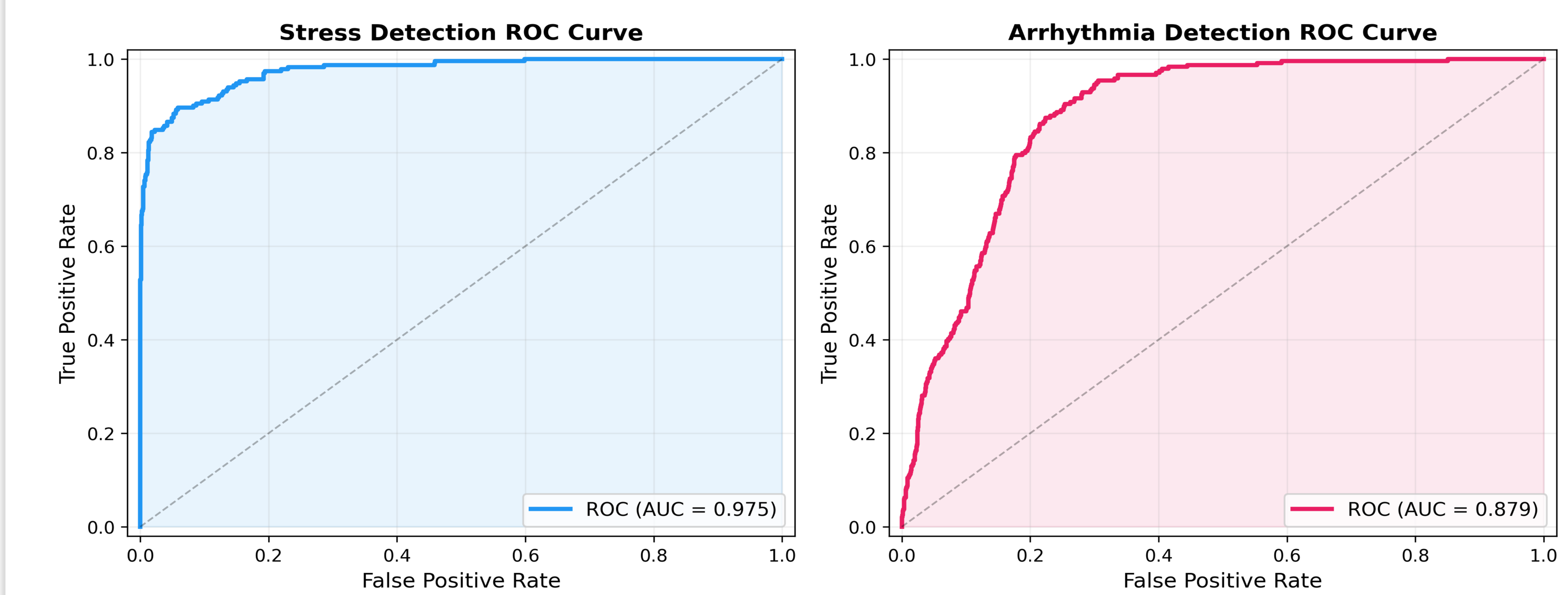


Figure 3. ROC curves: Stress (AUC 0.975) near-perfect; Arrhythmia (AUC 0.879) strong; Activity (AUC 0.731) limited by class imbalance

7 KEY FINDINGS & DISCUSSION

Classification Performance

- Stress AUC 0.975 — near-perfect discrimination across sensor modalities and subject populations
- Arrhythmia specificity 94.7%, false alert rate ~3% — conservative screening suitable for clinical use
- Activity accuracy 94.1%, but macro-F1 0.259 due to class imbalance (sedentary >80% of windows)

Deployment & Adaptation

- Full 450 KB FP32 model fits in PSRAM directly — no quantization, pruning, or distillation needed
- Domain shift from research-grade to consumer-grade sensors addressed by logit bias calibration
- On-device training: +8 KB flash, zero static RAM; A/B validation gate and safety lockout verified on live hardware

8 CONCLUSION & FUTURE WORK

Contributions

- First system combining multi-task edge inference + on-device adaptive training + 9 safety mechanisms on a sub-$15 microcontroller for wearable health monitoring
- Simultaneous monitoring of activity, stress, and arrhythmia enables detection of cross-domain physiological interactions (e.g., stress-induced arrhythmia patterns) invisible to single-task systems
- Complete bare-metal C++ implementation with zero dynamic allocation, no RTOS, no external ML libraries — suitable for safety-critical medical device firmware

Future Work

- Longitudinal clinical study with on-device adaptation generating first jointly-annotated activity–stress–arrhythmia corpus
- INT8 quantization to reduce inference latency below 1 second; incorporate UCI HAR for richer activity labels

9 ACKNOWLEDGEMENT

- The author thanks Dr. Shahnam Mirzaei for his guidance and mentorship throughout this research. This project was supported by the Thesis GRAD Center. Portions of this work have been accepted at IEEE ICHI 2026. Department of Computer Engineering, California State University, Northridge.

References: [1] Reiss et al., PPG-DaLiA: A Wrist-Worn Dataset for PPG-Based Heart Rate Estimation, 2019 [2] Schmidt et al., WESAD: Multimodal Dataset for Wearable Stress and Affect Detection, 2018 [3] Moody & Mark, The Impact of the MIT-BIH Arrhythmia Database, 2001 [4] Ren et al., TinyOL: TinyML with Online Learning on Microcontrollers, 2021 [5] Cai et al., TinyTL: Reduce Memory, Not Parameters, 2020 [6] Kwon et al., TinyTrain: Deep Neural Network Training at the Extreme Edge, 2023