ARIA — The Personalization Layer for Human-State Intelligence

Your stress score is 40% noise. We fixed the calibration. A drop-in API that turns raw wearable data into per-user arousal, stress, and affect scores after a one-time 5-minute calibration. Today: arousal from the wrist. Next: emotions, focus, fatigue, and any state your users can label.

The Problem

Over 500 million wearables ship the same one-size-fits-all stress model to every user. Resting heart rate, skin conductance, HRV, and tonic arousal vary significantly between healthy adults — a single population model cannot represent that variance, so users stop trusting the score and the feature quietly gets deprioritized. Competitors respond by training on larger wearable corpora; Google's LSM-2 was trained on **40M hours of wearable data from 60,000+ participants** and still reaches F1 around 0.68 on narrow clinical prediction tasks. The real bottleneck is not model capacity. It is calibration.

The Solution — Labels, Not Models

ARIA is the missing API between raw wearable signals (PPG, EDA, accelerometer) and real human-state intelligence. Each user runs a one-time structured calibration session; every subsequent prediction is scored against *their* baseline, not a population average. No cloud retraining. Per-user models are small enough to store on-device, so raw sensor data can stay on the user's wrist by design.

- Send a 60-second window of raw sensor data → receive `{ label, confidence, arousal, probabilities }`
- Calibrate once per user with ~5 minutes of labeled samples
- Per-user model stored on your backend, or — once our Wear OS runtime lands — directly on the watch

The product is the personalization layer, not the stress score. Layer 1 ships arousal (stressed / neutral / calm) from the wrist *today*. The same calibration architecture is designed to extend to emotions (L2, wrist + voice — roadmap) and user-defined states like focus, fatigue, and cognitive load (L3 — roadmap, not yet validated).

Current Evidence (Layer 1: arousal)

Setting	Accuracy
Daily life, 30 self-report labels (DAPPER, N=84) — what to expect in product use	**~56%**
Consumer watch, PPG + accelerometer (GalaxyPPG, N=23)	71.8%
Lab, zero-shot, no calibration (WESAD, N=15)	79.4%
Lab, controlled stress-induction protocol (TSST, 5-min calibration, WESAD, N=15) — *not representative of daily use*	89.4%

Balanced accuracy, leave-one-subject-out. **Read the ~56% field row as the in-product floor**, not the 89.4% lab row — DAPPER measures real users self-reporting during daily life and is not statistically distinguishable across architectures after Bonferroni correction. This convergence is the paper's central finding: **labels are the bottleneck, not model choice**. The 89.4% lab number comes from a controlled stress-induction protocol (TSST) and is what calibration achieves *in that setting* — a **+10 point lift** over zero-shot. To our knowledge this is the first systematic measurement of how much personalized data each architecture needs, which is the headroom product teams can build on.

Why ARIA

- **On-device by design** — per-user models are small enough that raw sensor data never has to leave the watch. Unlike large model providers, we don't need to see your users' heartbeats to make this work.
- **5-minute structured calibration** in the lab protocol we validated — phone-delivered calibration for in-app onboarding is the next engineering milestone, not a shipped claim
- **Cross-device validated** on research-grade wristband (Empatica E4-class) and consumer smartwatch (Samsung Galaxy Watch 5 via GalaxyPPG) — additional devices in roadmap

- **Extensible architecture** — the same calibration mechanism is intended to serve any state the user can label

Status

- **Methodology under submission** to ACM IMWUT (premier academic venue for wearable and ubiquitous computing), May 1, 2026
- **Working prototype API** — FastAPI with WebSocket streaming endpoint and per-user model store (localhost today; production-deployable)
- **Validated on 122 subjects** across three independent public datasets — WESAD (N=15, lab), DAPPER (N=84, field), GalaxyPPG (N=23, consumer hardware); 437 automated tests and explicit ML leakage audits

Next Step

If a personalization layer for human-state intelligence is on your 2026 roadmap, reply and we will share the working API, benchmark scripts, and sample request/response payloads by email. We can scope a pilot on your hardware once there is mutual fit.

Marco Accardi, CTO — Anecoica · info@anecoica.net · https://anecoica.us/