Into the Wild: Reliable Physiological Sensing with on-device Autoencoder-based Anomaly Detection

Kevin Penner, Felix Wittenfeld, Marc Hesse, Michael Thies
Cognitronics & Sensor Systems, Bielefeld University, Germany
{kpenner, fwittenf, mhesse, mthies}@techfak.uni-bielefeld.de

***Abstract*—Pushing physiological sensing into the wild: this work equips our ultra-low-power physiological BI-Vital sensor with an *int8*-quantized autoencoder that produces a real-time signal-quality index (SQI) for electrocardiograms (ECG) beside standard heart rate computations. The supervised model, trained on PhysioNet CinC2017 and three newly annotated BI-Vital single-lead ECG datasets, combines a compact encoder-decoder with a lightweight classifier head. On the STM32L476JE microcontroller from BI-Vital, the model uses 31 kB of RAM and 85 kB of flash memory, completes inference in 195 ms and still leaves room for parallel sensor services. Across eight train/test splits the approach generalizes well, reaching a F1 score of 0.988 while maintaining $2 \cdot 10^{-3}$ performance loss after quantization. By converting raw continuous ECG data into periodic heart rate and SQI recordings, storage space on the device is reduced by 99.6 %, demonstrating a practical way for reliable, data-efficient monitoring outside the laboratory.**

***Index Terms*—wireless body sensor, wearable, autoencoder, ecg anomaly, signal quality index (SQI), ultra-low-power microcontroller, TinyML, on-device inference**

I. Introduction

Physiological sensor technologies have advanced significantly, enabling extensive real-world monitoring applications in health, sports, and cognitive sciences. However, despite these advances, much of today's physiological knowledge is still rooted in laboratory-based research. While controlled experimental designs ensure the quality and reproducibility of data, they often do not address the complexity and variability of daily life [1]–[3]. Especially in affective and cognitive domains, limited embodiment, artificial stimuli, and restricted settings can lead to poor generalizability of results [4], [5]. To close this gap, data are increasingly being collected in the real world using wearable physiological sensors. These systems enable long-term and mobile monitoring, but also face a number of challenges. Capturing and storing raw signals, such as high-resolution electrocardiograms (ECG), is often impractical because continuous sampling, storage, and wireless transmission quickly exceed the capabilities of ultra-low-power (ULP) wearables [6]. A common solution is to compute data at a higher level of abstraction (e.g. heart rate) directly on-device. However, discarding the original raw signals makes it impossible to verify the reliability of the signals retrospectively and reduces the interpretability and scientific value of the derived metrics. Physiological monitoring therefore requires not only derived metrics but also a signal quality index (SQI) that quantifies their reliability [7], [8]. Manually defined rules or statistical SQI metrics often do not generalize well, whereas modern machine learning (ML) methods are not compatible with resource-limited edge hardware. This work proposes a quantized autoencoder (AE) with a lightweight classifier head that delivers real-time SQI estimation on the BI-Vital sensor [9]. Although demonstrated on single-lead ECG, the methodology is also applicable to other physiological modalities. Tests with the public CinC2017 dataset [10] and three own recorded datasets confirm robust generalization across devices. System measurements also show that the embedded model meets real-time requirements while significantly reducing flash and RAM usage. Continuous recording of raw signals can be replaced by compact recording of heart rate and SQI values, highlighting the practical benefits of on-device processing.

II. Related Works

Conventional SQIs based on fixed rules or simple statistics adapt poorly to different devices and activities [7], [8]. Therefore, unsupervised AEs have been used to derive data-driven SQIs, which provide more stable ratings under various noise conditions [11]. Further work extends the idea with adversarial AEs plus temporal convolutional neural networks for abnormal beat detection [12] and with a multi-scale masked AE [13]. Transfer learning studies also report faster convergence on small ECG sets, but the benefits of transfer learning disappear as the validation datasets grow [14]. On the hardware side, TinyML benchmarks demonstrate the cost of edge inference on heterogeneous platforms [6], [15]. Proof-of-concept ECG pipelines have been run on higher-power single-board computers (Jetson NX, Raspberry Pi) and microcontrollers (µC) such as the ESP8266, targeting denoising or arrhythmia classification rather than real-time SQI computing on ULP wearables [16], [17]. The present work addresses this gap by introducing a supervised AE quantized to *int8* that outputs SQI in real-time, explores depth-size trade-offs, observes strict memory budgets on the µC of the BI-Vital sensor, and generalizes from the public CinC2017 dataset to three self-labeled datasets. In addition, the SQI module is integrated into the BI-Vital firmware along with other sensor services as a TinyML node [9], unlike previous standalone demos, and enables quality feedback during multi-sensor acquisition.

III. Methods

A. Datasets and Preprocessing

Four datasets are used for the analysis: a subset of the public PhysioNet CinC2017 dataset and three self-acquired BI-

Vital datasets, each representing different real-world recording conditions [18]. The BV-Run dataset covers ECG recordings obtained during a university relay race and is characterized by frequently occurring motion artifacts related to the participants' movements. The BV-Cognitive dataset consists of ECG signals recorded during a controlled cognitive testing scenario focusing on reduced physical activity and minimal artifacts. BV-Children contains unsupervised recordings of children moving freely in a recreational environment. An overview of all datasets, including demographics, label distributions, and device-specific ECG parameters, is presented in Table I.

Dataset	**CinC2017 [10]**	**BV-Run**	**BV-Cognitive**	**BV-Children**
Participants	unknown	11 (9m, 2f), age 24–41	16 (14m, 2f), age 22–55	5 children, age <18
Year	2017	2023	2023	2024
Recordings	8528	11	16	55
Device	AliveCor	BI-Vital Sensor [9]		
Sampling Freq. [Hz]	300	200		
ADC Gain [counts/mV]	1000	1138		
ADC Res. [bit]	16	12		
Baseline [counts]	0	1470		
Labeled Windows	91,962	29,127	26,255	33,236
Signal	62,339	24,495	25,780	16,304
Noise	19,881	2,925	142	16,420
Unknown	9,742	1,707	333	512
Total valid Samples	Signal: 128,918 (76.6%), Noise: 39,397 (23.4%)			

TABLE I: Overview of datasets, demographics, device/ECG signal specifications and label distributions.

To ensure comparability across datasets, CinC2017 is downsampled to 200 Hz to match the native sampling rate of BI-Vital data. A fixed window size of 256 samples (1.28 s) ensures that most segments contain at least one full cardiac cycle, enabling the AE to learn and reconstruct characteristic ECG morphological features (P wave, QRS complex, T wave). This design avoids overlap or sliding windows, minimizing inference frequency and supporting efficient deployment in embedded systems.

Each window is manually labelled using a conservative noise-annotation scheme. Clearly corrupted segments are labeled as *Noise*, while ambiguous windows could optionally be marked as *Unknown* to support labeling consistency. These *Unknown* segments are subsequently excluded from the training dataset. To support consistent labeling, the interface displayed raw ECG traces, R-peaks (via NeuroKit2 [19]), and reference metrics such as kurtosis, pSQI, basSQI, and the Zhao2018 score. For CinC2017, original annotations were shown as context but not used directly. All signals are converted to mV units to ensure consistency between datasets:

$$\mathrm{ECG}_{\mathrm{mV}} = (\mathrm{ECG}_{\mathrm{raw}} - \mathrm{baseline}) \cdot {}^{1}\!/_{\mathrm{gain}_{\mathrm{ADC}}} \tag{1}$$

Conversion parameters are taken from the CinC2017 documentation and from the BI-Vital hardware specifications.

To evaluate model generalization, training and validation are performed for eight separate dataset splits. Each split is defined by specific combinations of training and validation datasets listed in Table II.

B. Autoencoder Design & Tuning

In this study, two convolutional AE model variants for ECG signal quality estimation are implemented (see Fig. 1), each using distinct strategies for training and inference.

Split ⟶	1	2	3	4	5	6	7	8
CinC2017	🟨	🟥		🟩	🟥	🟩	🟩	🟩
BV-Run	🟨		🟥	🟥	🟩	🟩	🟩	🟥
BV-Cognitive	🟨		🟥	🟥	🟩	🟩	🟥	🟩
BV-Children	🟨		🟥	🟥	🟩	🟥	🟩	🟩

TABLE II: Cross-validation configuration of training and validation datasets. Cells colored green indicate training data, red indicate validation data, white is unused. A yellow column means an internal 80 % train, 20 % validation split was used.

(1) Unsupervised AE: The unsupervised AE is trained exclusively on clean ECG segments. The objective is to learn a compact latent representation that accurately reconstructs clean signals. During inference, the reconstruction error $\mathcal{L}_{\mathrm{RE}}$ is used as an indicator of signal quality to separate clean from noisy ECG segments:

$$\text{MSE: } \mathcal{L}_{\mathrm{RE}} = {}^{1}\!/_{T} \sum_{t=1}^{256} (x_t - \tilde{x}_t)^2 \tag{2}$$

In addition, two alternative discriminator functions for mapping reconstruction error to quality scores are evaluated: logarithmic mean squared error (LOGMSE) and mean absolute deviation from MSE (MADMSE). Although trained unsupervised, the evaluation uses a supervised, threshold-based discriminator method to determine how effectively the reconstruction error separates clean and noisy segments.

(2) Supervised AE with Classifier Head: In this configuration, a dense classification head with softmax or sigmoid activation is appended to the output of the decoder. The model is trained end-to-end using both signal and noise samples. Two loss functions are evaluated: binary cross-entropy (BCE) and sparse categorical cross-entropy (SCCE), selected based on label representation and output format:

$$\text{BCE: } \mathcal{L}_{\mathrm{CL}} = -\left[y \cdot \log(\tilde{y}) + (1-y) \cdot \log(1-\tilde{y})\right] \tag{3}$$

$$\text{SCCE: } \mathcal{L}_{\mathrm{CL}} = -\log\left(\tilde{y}_y\right) \tag{4}$$

Architectural Parameters: The number of Conv1D layers n in the encoder and decoder is systematically varied between 2 and 7, exploring trade-offs between model complexity, performance, and suitability for µC deployment. All architectures use a fixed latent space dimension of 16 and an input window size of 256 ECG samples. Optimization is performed using the Adam optimizer with a learning rate of 10^{-4}, batch size of 32, and training up to 200 epochs, with early stopping after 20 epochs of no improvement.

Cross-Domain Evaluation: All models are evaluated across multiple training-validation splits covering different datasets and domains (see Table II) to test generalizability across devices and recording conditions.

Model Quantization: To evaluate deployment realizability, all trained models are converted to *.tflite* format using post-training quantization in four different quantization configurations. Each variant is evaluated based on model size and accuracy degradation.

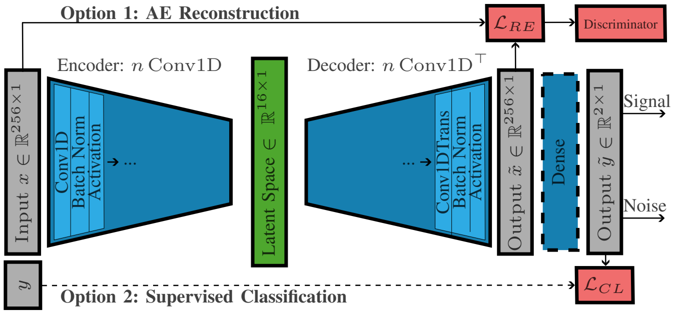
Option 1: AE Reconstruction
$\mathcal{L}_{RE}$
Discriminator
Encoder: n Conv1D
Decoder: n Conv1D$^{\top}$
Input $x \in \mathbb{R}^{256\times1}$
Conv1D
Batch Norm
Activation
→ ...
Latent Space $\in \mathbb{R}^{16\times1}$
... →
Conv1DTrans
Batch Norm
Activation
Output $\tilde{x} \in \mathbb{R}^{256\times1}$
Dense
Output $\tilde{y} \in \mathbb{R}^{2\times1}$
Signal
Noise
y
Option 2: Supervised Classification
$\mathcal{L}_{CL}$

Fig. 1: AE architecture variants: (1) Unsupervised AE utilizing reconstruction error for SQI estimation, (2) Supervised AE with an added classification head.

C. Evaluation Strategy

The performance of the model is evaluated quantitatively and qualitatively, with a focus on classification accuracy, generalizability and usability in embedded environments.

a) Classification Performance: Each trained model, including hyperparameter variants and quantized versions, is validated on the corresponding validation dataset. Evaluation metrics comprise classification accuracy, F1 score, and receiver operating characteristic (ROC) curves (including false positive rate, FPR, and true positive rate, TPR). These metrics facilitate a comparative analysis across different model configurations.

b) Threshold-Based SQI Evaluation: In addition to end-to-end models, the signal-to-noise separability is tested with conventional SQIs (Kurtosis, Skewness, IOR, pSQI, basSQI and Zhao2018) [7], [8], [11] using a threshold discriminator for comparison with learning-based methods.

c) Memory Footprint: To determine suitability for µC deployment, the memory usage of each quantized model variant is examined. Flash and RAM footprints are obtained via the PlatformIO build system. This analysis verifies whether each model fits within the constraints of embedded hardware.

d) Inference Runtime: Inference execution time is measured directly on the BI-Vital sensor. Other sensor services are disabled to isolate runtime performance, and each model variant runs as a standalone application. Inference should be completed within a window length of 1.28 s, which matches the ECG input to ensure real-time capability.

IV. Results and Discussion

Classification Performance and Benchmarks:

Fig. 2(a) presents the classification performance of various AE model configurations for ECG signal quality estimation, analyzing the influence of network depth, the inclusion of a classification head, and the choice of reconstruction-based discrimination strategy. All results refer to the Split 1 dataset, where both the training and validation sets come from the entire dataset pool. Models that include a classification layer are consistently better than models based only on reconstruction errors, particularly in terms of TPR and FPR. Although reconstruction-based models exceed the majority of traditional statistical baselines, skewness is competitive among unlearned methods. The Zhao2018 metric achieves high TPR but suffers from poor FPR, frequently misclassifying noisy signals as valid ECGs. This behavior limits its reliability under uncontrolled conditions.

Fig. 2(b) analyzes the memory footprint of the most promising model types (classification head with BCE or SCCE loss) after applying different levels of post-training quantization. As expected, deeper networks result in larger models. Quantization substantially reduces memory usage: model size drops by up to 69 % for deep architectures and by approx. 44 % for compact configurations. The trade-off in classification performance remains insignificant, with the reduction in F1 score limited to below $2 \cdot 10^{-3}$.

Fig. 2(c) evaluates the smallest model *Very shallow*—$2 \cdot conv\_blocks$ $(kernel_{size} = 9, strides = 4, filters = 12, padding = "same")$ with a classification head, BCE loss and int8 quantization — across the dataset splits. The results confirm strong generalization capabilities across diverse validation splits. The weakest performance, with an F1 score of 0.956, occurs when training on all BI-Vital datasets and validating on the CinC2017 dataset. Vice versa, validation performance on BI-Vital data remains strong throughout, with F1 scores of 0.991, supporting applicability across single-lead ECG systems, provided signals are normalized to mV units.

Memory Footprint and Inference Time:

The deployability of the quantized models is tested on the BI-Vital hardware, which integrates an STM32L476JE µC with 96 kB of RAM and 512 kB flash memory. Inference is performed using the TensorFlow Lite Micro (TFLM) library as the runtime backend. With a requirement of 31.14 / 30.19 kB RAM (*int8* / *float32*) and 84.57 / 93.77 kB flash for the *very shallow* model, the realization is suitable for execution on the BI-Vital even alongside the existing software components. With an execution time of 194.6 / 175.3 ms, the real-time requirement of 1.28 s is also clearly met, ensuring practical use while other system tasks run. In terms of long-term data recording, the conversion from raw ECG logging to abstraction-based signal representation brings considerable storage benefits. A raw ECG at 200 Hz stored as *int16* yields 400 $^{bytes}/_{s}$. By replacing each 256-sample window (512 bytes) with two *uint8* values (heart rate & SQI), the data stream is compressed at a ratio of 256 (99.6 %) to 1.56 $^{bytes}/_{s}$, enabling long-term on-board storage and lower bandwidth for low-power wireless transmission.

V. Conclusion

In this paper, a lightweight quantized AE-based model for real-time ECG signal quality estimation on embedded hardware with ULP requirements is presented. Supervised models with classification layers show superior performance over reconstruction-based approaches and traditional SQIs. Quantization to *int8* format reduce the model's memory footprint without significant loss of accuracy, facilitating its use on resource-constrained µCs. The model shows strong

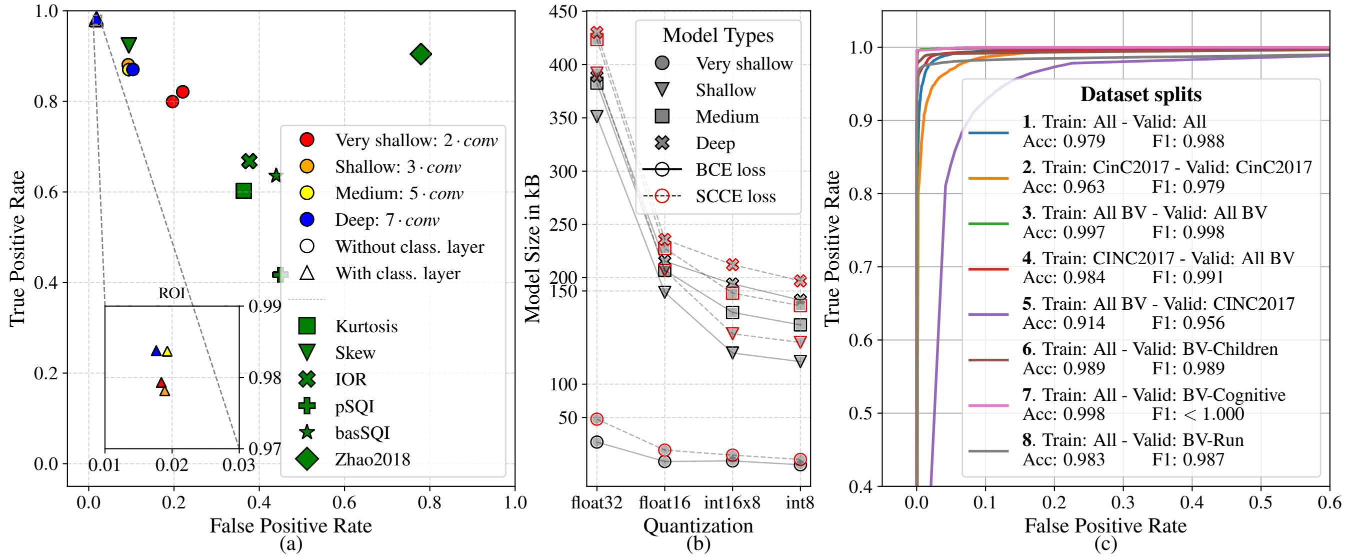

Fig. 2: Evaluation of classification performance, model footprint, and cross-domain generalizability. (a) Classification results for different model configurations and conventional SQI baselines. (b) Model size across quantization levels and depths. (c) ROC curves for the most compact model *Very shallow* across eight dataset splits.

generalization across different datasets and recording conditions. Furthermore, by replacing continuous raw ECG data recording with periodic heart rate and SQI measurements, data compression is achieved, improving storage efficiency in long-term monitoring. The resulting dataset, which includes approx. 88,600 labeled ECG windows from the BI-Vital sensor [18] and 92,000 relabeled windows from the CinC2017 dataset, provides a valuable resource for future research in physiological signal analysis and model development.

Real-time detection of measurement anomalies on-device opens up possibilities for applications such as user feedback on inappropriate use of the device and serving as a real-time wearable assistant. Extending the model to multi-channel ECG or other physiological signals such as PPG and EMG could extend its applicability. Integrating the detection of medical anomalies with the validation of measurement errors could improve diagnostic capabilities. In addition, running the model on ULP AI accelerators such as the Syntiant NDP120 or GreenWaves GAP9 promises further improvements in inference efficiency and energy consumption [20], increasing the practical utility of embedded biosignal monitoring systems.

VI. Acknowledgments

We would like to thank the Core Facility for Neuroscience of Self-Regulation (CNSR) at Heidelberg University for providing the BV - Children ECG data set and for the cooperation in the field of psychophysiological real-world methodology.

References

[1] M. Chang *et al.*, “Ecological validity in exercise neuroscience research: A systematic investigation,” *European Journal of Neuroscience*, vol. 55, pp. 487–509, 2022.

[2] S. G. Shamay-Tsoory *et al.*, “Real-life neuroscience: An ecological approach to brain and behavior research,” *Perspectives on Psychological Science*, vol. 14, pp. 841–859, 2019.

[3] N. Chaytor *et al.*, “The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills,” *Neuropsychology Review*, vol. 13, pp. 181–197, 2003.

[4] S. Dikker *et al.*, “Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom,” *Current Biology*, vol. 27, pp. 1375–1380, 2017.

[5] J. E. Crum II, “Future applications of real-world neuroimaging to clinical psychology,” *Psychological Reports*, vol. 124, pp. 2403–2426, 2020.

[6] C. R. Banbury *et al.*, “Benchmarking TinyML systems: Challenges and direction,” *CoRR*, 2020.

[7] Z. Zhao *et al.*, “SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation,” *Frontiers in Physiology*, vol. 9, 2018.

[8] F. Liu *et al.*, “An overview of signal quality indices on dynamic ECG signal quality assessment,” *Feature Engineering and Computational Intelligence in ECG Monitoring*, pp. 33–54, 2020.

[9] K. Penner *et al.*, “TinyML optimization for activity classification on the resource-constrained body sensor BI-Vital,” *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 2023.

[10] G. D. Clifford *et al.*, “AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017,” *2017 Computing in Cardiology (CinC)*, pp. 1–4, 2017.

[11] N. Seeuws *et al.*, “Electrocardiogram quality assessment using unsupervised deep learning,” *IEEE Transactions on Biomedical Engineering*, vol. 69(2), pp. 882–893, 2022.

[12] L. Shan *et al.*, “Abnormal ECG detection based on an adversarial autoencoder,” *Frontiers in Physiology*, vol. 13, 2022.

[13] Y. Zhou *et al.*, “Multi-scale masked autoencoder for electrocardiogram anomaly detection,” 2025.

[14] C. Nguyen *et al.*, “Transfer learning in ECG diagnosis: Is it effective?” *PLOS ONE*, vol. 20, 2025.

[15] D. Samakovlis *et al.*, “BiomedBench: A benchmark suite of TinyML biomedical applications for low-power wearables,” 2024.

[16] Y. Jhang *et al.*, “Integration design of portable ECG signal acquisition with deep-learning based electrode motion artifact removal on an embedded system,” *IEEE Access*, vol. 10, pp. 57 555–57 564, 2022.

[17] M. Hizem *et al.*, “Reliable ECG anomaly detection on edge devices for internet of medical things applications,” *Sensors*, vol. 25, 2025.

[18] K. Penner *et al.*, “will be published after the review,” 2025.

[19] D. Makowski *et al.*, “NeuroKit2: A python toolbox for neurophysiological signal processing,” *Behavior Research Methods*, vol. 53, 2021.

[20] M. Kaiser *et al.*, “VEDLIoT: Very efficient deep learning in IoT,” *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 963–968, 2022.