A Compact, Uncertainty-Aware mLSTM Model for Real-Time ECG Delineation on Edge Devices

David Lung, Philipp Heute, Manfred Marx, Thomas Schlöglhofer, Theodor Abart, Francesco Moscato, Julia Riebandt, Daniel Zimpfer, and Max Haberbusch, *Member, IEEE*

This work has been submitted to the IEEE for possible publication.
Copyright may be transferred without notice, after which this version
may no longer be accessible.

***Abstract*— Accurate electrocardiogram (ECG) waveform delineation underpins cardiac diagnosis and screening, yet the most accurate deep-learning models are too large for resource-constrained wearable and implantable systems. We present a compact, uncertainty-aware bidirectional matrix-memory LSTM (Bi-mLSTM) for per-sample P, QRS, and T delineation with only 2,522 parameters. The model operates on beat windows placed by automatic R-peak detection. It was selected and trained on the automatic annotations of the PhysioNet QT Database, frozen, and evaluated against expert annotations of the Lobachevsky University Database (LUDB, 200 subjects) and a clinical cohort of 25 patients with left ventricular assist devices (LVAD-DB). The tolerance-aware macro-F1 was 0.919 on LUDB and 0.875 on LVAD-DB (strict 0.850 and 0.761), the expected calibration error of the raw confidence was 0.054 and 0.041, and low confidence flagged the model's own errors. Under tolerance-aware scoring, the model exceeded three published architectures with 0.3 to 1.4 million parameters, trained with the same windowing and procedure, on LUDB and matched the strongest of them on LVAD-DB with 119 times fewer parameters. Per record, model-derived QRS duration and QT exceeded the expert values by 12.7 and 11.3 ms on LUDB, which supports trend monitoring rather than replacement of expert measurement. On an Arduino Portenta H7 microcontroller, a separately trained earlier version of the same architecture delineated a live recording in real time without quantization, with 128.6 ms of processing per beat and 22 kB of memory. Accurate, calibrated, and uncertainty-aware ECG delineation is therefore feasible in real time on resource-constrained devices.**

***Index Terms*— Electrocardiogram waveform analysis, real-time signal processing, recurrent neural networks, wearable medical devices, predictive uncertainty**

This work was supported by the Austrian Research Promotion Agency (FFG) under the CARDISENSE project (grant FO999922422). *(Corresponding author: Max Haberbusch.)*

David Lung, Philipp Heute, Theodor Abart, Francesco Moscato, and Max Haberbusch are with the Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, 1090 Vienna, Austria. Theodor Abart, Francesco Moscato, and Max Haberbusch are also with the Ludwig Boltzmann Institute for Cardiovascular Research, Vienna, Austria. Francesco Moscato is also with the Austrian Cluster for Tissue Regeneration, Vienna, Austria. Manfred Marx is with the Department of Pediatric Cardiology, Medical University of Vienna, Vienna, Austria. Thomas Schlöglhofer, Julia Riebandt, and Daniel Zimpfer are with the Department of Cardiac Surgery, Medical University of Vienna, Vienna, Austria; Thomas Schlöglhofer is also with the Center for Medical Physics and Biomedical Engineering and the Ludwig Boltzmann Institute for Cardiovascular Research, Vienna, Austria. (e-mail: max.haberbusch@meduniwien.ac.at).

I. Introduction

Electrocardiogram (ECG) delineation, the localization of the P wave, QRS complex, and T wave and their temporal boundaries, is foundational to clinical diagnostics and automated cardiac analysis [1]–[3]. It supports arrhythmia classification, ischemia detection, and the evaluation of structural heart disease [4]. Rule-based methods built on hand-engineered features often struggle with morphological variability, noise, and pathological alterations [3], [4]. Delineating on the sensing device itself would make interval measures such as PR, QRS, and QT available continuously and in real time, without streaming raw signals off the body.

Deep learning with convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers has substantially advanced delineation [5]–[16], and hybrid networks combining CNNs with long short-term memory (LSTM) approach human performance on standard datasets. Many of these models, including DENS-ECG [5] and deep U-Net variants [7], exceed one million parameters. This obstructs their use in resource-constrained environments such as mobile, wearable, and implantable devices, where real-time inference, energy efficiency, and small memory footprints are critical [17]. Recent edge-oriented networks address beat classification rather than delineation. For example, Rahman and Morshed [18] ran a feature-based LSTM beat classifier on a custom wearable with a 64-MHz Arm Cortex-M4, reaching 20-ms inference with 6.7 kB of peak RAM on MIT-BIH data, and Sarhaddi et al. [19] proposed spiking recurrent classifiers of 19.6k to 41.4k parameters for MIT-BIH and PTB-XL whose computational cost was estimated in software, without embedded deployment. Such classifiers are trained for one task and return a label without the physiological measurements behind it. Delineation instead yields the clinical intervals that clinicians already use, namely the P-wave duration, the PR interval, the QRS duration, and the QT interval with its heart-rate correction (QTc). These intervals are interpretable and trusted, and simple downstream models can reuse them for many diagnostic tasks.

The extended LSTM (xLSTM) [20] introduces a scalar-memory (sLSTM) and a matrix-memory (mLSTM) variant of the LSTM cell with decoupled, parallelizable memory operations, increasing modeling capacity without a corresponding rise in parameter count. xLSTM has shown benefits in natural-language and sequential-data tasks and has been applied to ECG classification and signal synthesis [21], [22]. xECG, an xLSTM-based ECG foundation model with about 57 million parameters, performs strongly across a broad benchmark of ECG tasks [23]. xLSTM has not yet been applied to fine-grained waveform delineation, nor at a scale that fits a microcontroller.

In this work, we propose the first mLSTM-based model for sample-wise ECG delineation. The bidirectional model comprises only 2,522 parameters and assigns each ECG sample to the P-wave, QRS-complex, T-wave, or no-wave class. Its performance is validated on a public benchmark, on a clinical cohort of patients with left ventricular assist devices (LVADs), and in real-time operation on a wearable device (Fig. 1a), and is compared with that of three established delineation architectures. In addition, clinical intervals derived from the model output are evaluated against those derived from expert annotations.

II. Materials and Methods

A. Datasets and Annotation

Models were trained and validated on the PhysioNet QT Database (QTDB) [4], [24], which comprises 105 two-lead records of 15 min. The training labels were the automatic single-lead wave boundaries distributed with QTDB (annotation files pu0 for the first and pu1 for the second lead), which cover every beat of each record. The manual QTDB annotations, which cover about 30 beats per record, were not used. Both leads were used as independent single-lead signals. The model therefore required no manual delineation of its training data, and every external evaluation compares it with expert manual annotations.

External evaluation used two datasets that were not used for training or model selection. The Lobachevsky University Database (LUDB, 200 subjects) [4], [25], [26] provides manually annotated 12-lead ECGs. LVAD-DB comprises 28 continuous 24-hour 3-lead ambulatory Holter recordings (250 Hz) from 25 patients with advanced heart failure on HeartMate 3 LVAD support (three patients contributed two recordings on separate dates). In these recordings, segments of approximately 5–10 s were selected for annotation at roughly hourly intervals, spanning diurnal and activity-related changes in rhythm and morphology. No patient was in atrial fibrillation and no pump-related artifacts were present in the analyzed segments. LVAD-DB was annotated sample by sample by a cardiologist (23 patients, 26 recordings) and by a biomedical engineer, who annotated five continuous 5-min segments from recordings of the same cohort. Each expert was blinded to the other's labels and to the model outputs. Both experts also annotated the on-device pilot recording. Agreement is pooled across raters unless stated otherwise. The LVAD-DB study protocol was approved by the Ethics Committee of the Medical University of Vienna (identification number EK1251/2020), and written informed consent was obtained from all participants, including the healthy volunteer of the on-device pilot.

B. Preprocessing and Beat Windowing

Signals were resampled to 250 Hz where necessary and band-pass filtered (0.5–40 Hz) (Fig. 1b). Beat windows were placed by an automatic Pan–Tompkins-style R-peak detector with a 5–15 Hz detection band on every dataset. Each detected R peak was converted to a QRS-onset anchor by subtracting one constant of 15 samples (60 ms), the median of the per-record delays from R peak to QRS onset in the 83 QTDB training records. The same constant was applied unchanged to every external dataset, so that no external annotation entered the windowing. A window starts at one anchor and extends to the next, so that it spans one cardiac cycle comprising the QRS complex, the T wave, and the following P wave. Intervals longer than 300 samples (1.2 s) were excluded, and each remaining window was right-padded with zeros to 300 samples and z-score normalized. Every sample is labelled as P wave, QRS complex, T wave, or No-Wave (NW), the last covering the isoelectric and inter-wave segments outside the three waves.

Training and evaluation used the same detector and window geometry, so every reported metric includes R-peak detection errors and the beat-to-beat jitter between the detected anchor and the true QRS onset. Offline, the detector searched only within the annotated stretches of each record, padded by 0.4 s on either side. Windows started by detections in this padding or by false detections were scored against the reference labels, whereas a missed beat produced an interval longer than 300 samples and therefore entered the results only through the detector sensitivity.

C. Model Architecture and Training

The model is built from the mLSTM block of the xLSTM framework [20] (Fig. 1c). A linear projection of the input feeds a stack of mLSTM blocks. Each block projects its input up, applies a depthwise causal convolution before the query and key projections, updates the matrix memory, and projects the gated output back down. There is no separate convolutional stem. In the bidirectional model (Bi-mLSTM) the same stack is applied to the input and to its time reversal with shared weights, the two layer-normalized outputs are concatenated, and a linear layer with a per-sample softmax yields the class probabilities. The selected configuration has 2,522 parameters (embedding dimension 8, three blocks, one head, convolution kernel 11), and its full layer structure, including the placement of every normalization, is given in Supplementary Section S3.

The architecture was selected by multi-objective Bayesian optimization (Optuna, 60 trials) that jointly maximized QTDB validation macro-F1 and minimized parameter count (search space and representative configurations in Supplementary Tables S1 and S2). The model at the knee of the Pareto front of QTDB accuracy versus parameter count was chosen. To keep the search inexpensive, the search and the architecture ablation

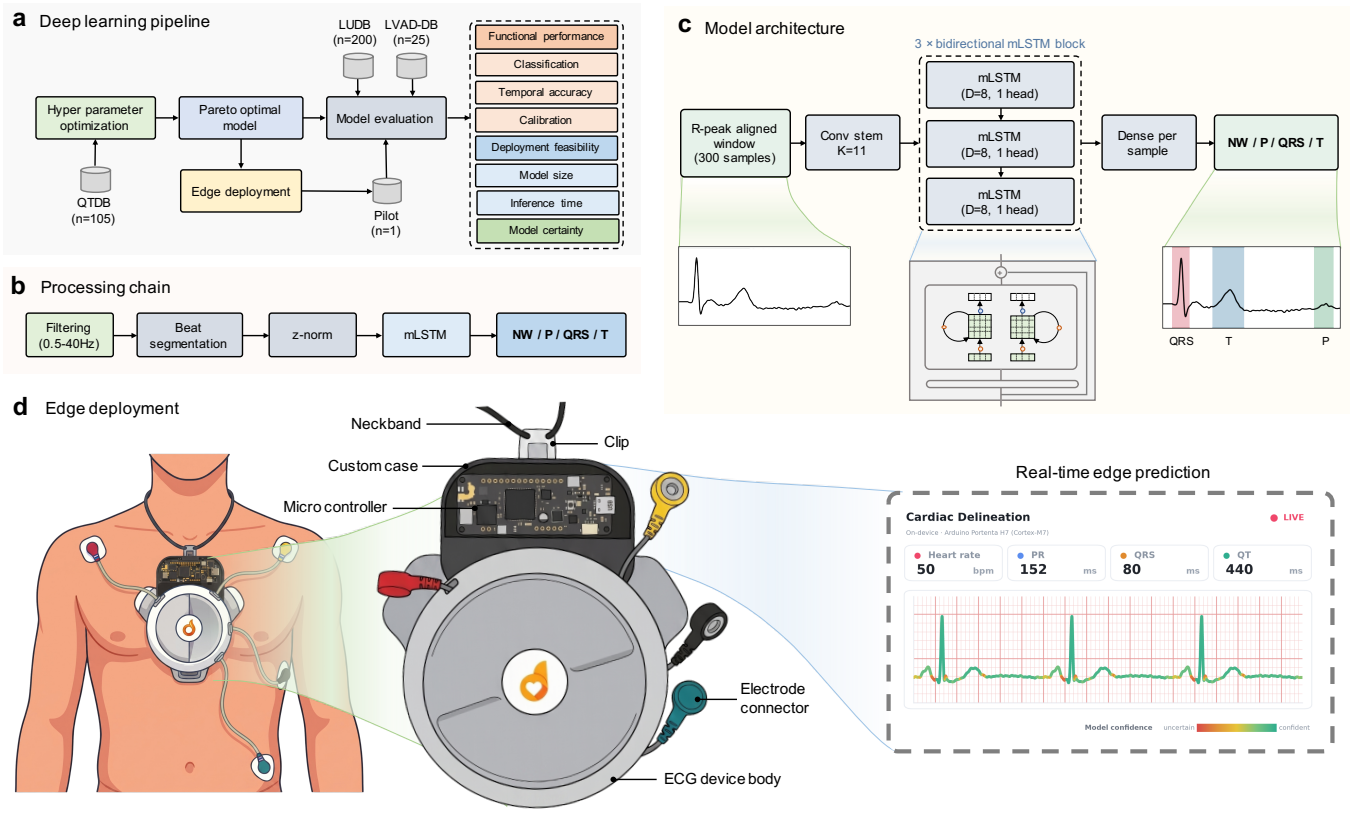


Fig. 1. System overview. **(a)** Study pipeline. Multi-objective Bayesian optimization on QTDB ($n = 105$ records) yields a Pareto-optimal model, which is frozen and evaluated on the external LUDB ($n = 200$) and LVAD-DB ($n = 25$) and, after edge deployment, in the on-device pilot ($n = 1$). **(b)** Processing chain with band-pass filtering (0.5–40 Hz), beat windows anchored at automatically detected R peaks, z-normalization, and per-sample classification into No-Wave (NW), P, QRS, and T. **(c)** Architecture with a linear input projection, three mLSTM blocks with internal depthwise convolution (kernel size 11, embedding dimension 8, one head) applied in both directions with shared weights, and a per-sample classifier (2,522 parameters). The inset shows a delineated beat. **(d)** Edge deployment. A portable ECG monitor streams one lead via BLE to an Arduino Portenta H7 (Cortex-M7) running the full pipeline, and the app colors the waveform by confidence (teal for confident, red for uncertain samples) and shows live HR, PR, QRS, and QT (Fig. 4).

placed windows at the annotated QRS onsets rather than at detected R peaks. The architecture and the final checkpoint were selected on QTDB alone and were frozen before any external evaluation. The external scores of the search configurations and of the ablation variants reported below were computed afterwards and are post-selection analyses that did not inform any design choice.

The final model was trained on QTDB beat windows placed by the R-peak detector, with each anchor set at the detected R peak minus the R-to-onset delay of its record. Windows without a labelled QRS sample were discarded. Record sel33 (median RR interval 1.7 s) contributed no window of at most 300 samples, so 104 records yielded 202,155 clean windows. Augmentation made the model robust to errors in window placement. Alongside the clean copy of every beat, three copies were cut at anchors displaced by Gaussian offsets (standard deviation 40 ms, clipped to ± 60 ms), which gave 806,613 training windows in total. Time stretching ($\pm 10\,\%$), sinusoidal baseline wander (0.1–0.5 Hz, amplitude up to 0.3 z-units), and Gaussian noise (standard deviation up to 0.15 z-units) were applied on the fly to the non-padded region (effect of augmentation in Supplementary Table S3). The records were split 80/20 into 83 training records (633,306 windows) and 21 validation records (173,307 windows), with both leads and all copies of a record in the same partition. The three QTDB subjects documented to have contributed two records each were all in the training partition, so no subject appeared in both partitions. The loss was categorical cross-entropy weighted by inverse class frequency, with the NW weight doubled and label smoothing ($\epsilon = 0.05$). Optimization used Adam with warm-up and cosine annealing, batch size 256, and early stopping on the validation tolerance-aware macro-F1. All hyperparameters and the implementation of the detector and the firmware are detailed in Supplementary Section S3.

D. Evaluation

Models were tested on LUDB and LVAD-DB without adaptation, on 35,558 and 5,205 beat windows, respectively. Tolerance-aware scores follow Peimankar et al. [5]. Within each window, a sample annotated as P, QRS, or T counts as correctly predicted if the model assigned the same class to at least one sample within ± 52, ± 48, or ± 124 ms (± 13, ± 12, and ± 31 samples at 250 Hz). Only predictions at annotated wave samples are relabelled, and only towards the annotation, so tolerance matching forgives under-segmentation and small boundary shifts but never over-segmentation into NW. Strict

sample-wise scores, without this relabelling, are reported alongside (per-class values in Supplementary Table S4). We report class-wise and macro-averaged F1, precision, recall, Cohen's κ, and area under the receiver operating characteristic curve (AUC). Strict F1, precision, and recall compare the argmax prediction with the expert labels. Tolerance-aware F1, precision, and recall, as well as κ, compare the tolerance-matched prediction with the same expert labels. AUC and the expected calibration error (ECE) are computed from the predicted probabilities against the original labels (binning in Supplementary Section S9). Calibration was assessed with the ECE of the raw confidence and with reliability diagrams. For post-hoc temperature scaling, a single scalar T was fitted by minimizing the negative log-likelihood on the QTDB validation records and applied unchanged to both external sets. Per-sample reliability was quantified by the predictive confidence (maximum softmax probability) and the normalized predictive entropy. R-peak detection was scored against a reference R peak placed at the largest deflection within each annotated QRS complex, with a 150-ms matching tolerance and only detections inside annotated stretches counted. Because the detector searched only the annotated stretches, its positive predictive value (PPV) does not include false detections in unannotated signal.

Boundary timing was assessed by Bland–Altman analysis on the detector-placed windows. Onsets and offsets were taken from the raw argmax prediction, and each expert boundary was matched to the nearest predicted boundary of the same class within the class tolerance. Unmatched boundaries were excluded, so the timing statistics are conditional on a match, and the matched fraction of the expert boundaries is reported for every boundary type.

Clinical intervals were derived per beat with identical rules from the predicted and the expert labels, which were stitched across contiguous windows into continuous label streams so that waves spanning a window boundary were measured intact. Segments touching the edge of a contiguous stretch were not used. Same-class segments separated by at most 20 ms were merged and segments shorter than 12 ms were ignored. For each beat, the QRS complex nearest to the window anchor, the preceding P wave, and the following T wave were identified. P-wave duration, PR interval, QRS duration, and QT interval were computed from their onsets and offsets, and QT was corrected for heart rate with the formulas of Bazett (QTcB) and Fridericia (QTcF) using the preceding RR interval. An interval was compared only when every wave it requires was found in both sources, and this fraction of beats is reported as the interval availability. Agreement was summarized per beat by the bias, standard deviation (SD), and mean absolute error (MAE), and per record by the difference between the model and expert medians of records with at least three measurable beats. In LUDB records in sinus rhythm without pacing, each record median was classified as within or outside the normative range (P-wave duration up to 120 ms, PR 120–200 ms, QRS up to 120 ms, QTc below 450 ms in men and 460 ms in women), and agreement between model and expert was quantified with Cohen's κ.

Confidence intervals were obtained by a patient-level bootstrap with 2,000 resamples and 95% percentile intervals (Fig. 2d). The resampling unit was the subject on LUDB (199 subjects with detector windows) and the patient on LVAD-DB, where the windows of both raters for a patient were kept together. Two of the engineer's recordings could not be linked to a patient code and were treated as separate resampling units. Resampling whole patients preserves the dependence between the beats, leads, recordings, and annotations of each patient. Differences to each baseline were tested with a paired bootstrap in which both models were scored on the same resampled patients. Two-sided bootstrap p-values were corrected with the Holm method over the three baselines and two datasets, and a difference was deemed significant when the adjusted p-value was below 0.05.

E. Baselines

The model was compared with three published delineation architectures with open implementations, the CNN–BiLSTM of Peimankar et al. [5] (1,419k parameters), the CNN of Jiménez-Pérez et al. [7] (1,296k), and the CNN–BiLSTM of Liu et al. [11] (301k). Nurmaini et al. [6] was omitted because no open implementation was available. The three architectures were re-implemented and trained from scratch with the same windowing, augmentation, loss, optimizer, schedule, batch size, data split, and stopping rule as our model, so that only architecture and size differ. In addition, each baseline was run in its native pipeline, using its original checkpoint on unanchored 1.2-s sliding windows. Because checkpoint, training data, and preprocessing differ there as well as the windowing, this comparison serves as descriptive context only.

F. Embedded Deployment and Pilot

The pipeline runs entirely on the single Arm Cortex-M7 core of an Arduino Portenta H7 (Fig. 1d), housed in a custom 3-D-printed enclosure that clips onto a D-Heart® (d-heartcare, Italy) portable ECG monitor worn on the chest at lead D2. Acting as a BLE central, the board receives single-lead ECG at 160 Hz from the CE-marked D-Heart monitor and buffers it in a 10-s ring buffer. It processes the signal in a heartbeat-paced loop that mirrors the offline pipeline, with polyphase resampling to 250 Hz, zero-phase Butterworth band-pass filtering (0.5–40 Hz), global z-normalization over a 5-s buffer, and on-device Pan–Tompkins-style R-peak detection that places one 300-sample window per beat. Each beat is classified as soon as the next R peak has been detected, from a 1.2-s window buffer, because the bidirectional model needs the complete cycle. Windows longer than 300 samples are truncated to their first 300 samples on the device, whereas they were excluded offline. On the device the windows are taken from the buffer-normalized signal, whereas offline each window was z-normalized individually. The 2,522-parameter model is implemented as hand-written C/C++ firmware under the Arduino/Mbed-OS RTOS and runs in 32-bit floating point, without quantization, hardware-specific acceleration, or an external inference runtime. Separate RTOS contexts for BLE reception, inference, and logging ensure that live reception is never starved during inference. Per-sample labels are stored

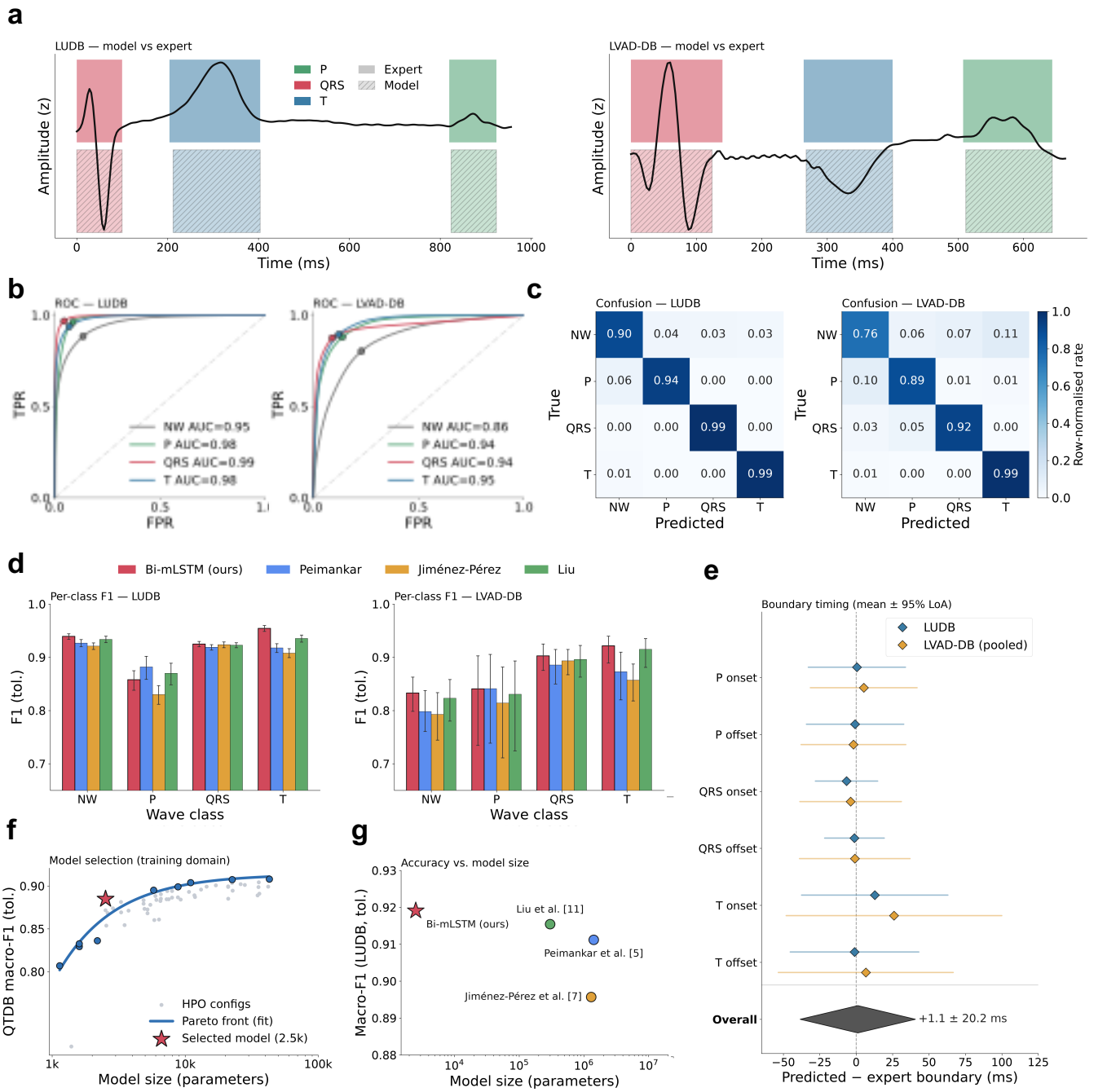


Fig. 2. Diagnostic performance on LUDB and LVAD-DB (pooled across raters). **(a)** Example beats with the highest model–expert agreement (per-sample accuracy 99.2% LUDB, 97.0% LVAD-DB), with expert annotation (upper band) and model prediction (lower band, hatched). **(b)** Per-class ROC curves with Youden points and AUC. **(c)** Row-normalized tolerance-aware confusion matrices. **(d)** Per-class F1 of the proposed model and the three baselines trained with the same windowing and procedure (Table III). Error bars show 95% patient-level bootstrap CIs. **(e)** Bland–Altman boundary timing (model minus expert, bias and 95% limits of agreement) at the onsets and offsets of the P wave, QRS complex, and T wave. Boundaries were taken from the raw prediction and matched to the nearest expert boundary of the same class within the class tolerance, so the statistics are conditional on a match. *Overall* pools all matched boundaries of both datasets (bias $+\mathbf{1.1} \pm \mathbf{20.2}$ ms). **(f)** QTDB model selection, showing macro-F1 against parameter count for the 60 search configurations, the Pareto front, and the selected model (star). Search-time scores were obtained on windows placed at the annotated QRS onsets. **(g)** LUDB macro-F1 against parameter count. Circles show the three baselines trained with beat-anchored windows (Table III) and the star the proposed model.

TABLE I
AGREEMENT AND CALIBRATION OF THE SELECTED MODEL ON THE EXTERNAL DATASETS (LVAD-DB POOLED ACROSS RATERS, 95% PATIENT-LEVEL BOOTSTRAP CIS). MACRO-F1 TOLERANCE-AWARE AND STRICT. ECE OF THE RAW CONFIDENCE AGAINST THE ORIGINAL LABELS.

Dataset	Macro-F1, tol. (95% CI)	Macro-F1, strict (95% CI)	Cohen's κ (95% CI)	ECE (95% CI)	Macro AUC
LUDB	0.919 (0.911–0.926)	0.850 (0.841–0.859)	0.891 (0.881–0.899)	0.054 (0.050–0.058)	0.977
LVAD-DB	0.875 (0.830–0.902)	0.761 (0.708–0.798)	0.832 (0.779–0.866)	0.041 (0.018–0.070)	0.923

in the board's QSPI flash or streamed to a host. For the N=1 pilot, a single healthy adult investigator followed a structured rest–walk–rest protocol. The device logged its per-sample predictions with per-beat latency and memory telemetry, and this logged output was scored with the same tolerance-aware pipeline as the offline evaluation.

III. RESULTS

A. Model Selection

The 60 search configurations spanned 1,140 to 42,980 parameters (median 6,576) with tightly clustered QTDB validation macro-F1 (median 0.885). QTDB macro-F1 saturated beyond approximately 5,000 parameters (Fig. 2f). The selected model is strictly Pareto-optimal and sits at the knee of the front. Compared with the next-smaller Pareto-optimal configuration it gains $+0.049$ QTDB macro-F1 for only $+334$ parameters, whereas an order of magnitude more parameters buys less than 0.03. In the post-selection analysis, external macro-F1 correlated strongly with QTDB macro-F1 across the search configurations (LUDB $r = 0.95$, LVAD-DB $r = 0.87$, both $p < 0.001$; Supplementary Fig. S1). A post-selection ablation showed diminishing returns in accuracy per parameter (Supplementary Table S5). Adding an sLSTM block increased the parameter count by 55% but raised macro-F1 by only $+0.030$ on LUDB and $+0.008$ on LVAD-DB, consistent with the mLSTM-only design selected on QTDB. A kernel of 41 raised LUDB macro-F1 by $+0.046$, a second attention head reduced LVAD-DB macro-F1 by 0.021, and a two-block stack reduced LUDB macro-F1 by 0.031.

B. External Agreement and Boundary Timing

The selected model reached a validation macro-F1 of 0.937 on QTDB. On the external datasets its tolerance-aware macro-F1 was 0.919 (95% CI 0.911–0.926) on LUDB and 0.875 (95% CI 0.830–0.902) on LVAD-DB, and its strict macro-F1 was 0.850 (0.841–0.859) and 0.761 (0.708–0.798) (Table I). QRS delineation was robust on both datasets (F1 0.925 and 0.903, AUC 0.993 and 0.936), whereas P and T waves were more domain-sensitive (Fig. 2a–d). Residual errors concentrated between NW and the low-amplitude P and T waves (Fig. 2c), and bootstrap intervals were wider on the smaller LVAD-DB cohort (Fig. 2d).

Agreement was higher with the biomedical engineer than with the cardiologist (macro-F1 0.938 vs. 0.824, κ 0.918 vs. 0.765; Supplementary Table S6 and Fig. S2). In six recordings the cardiologist labelled wide, biphasic ventricular complexes as a single QRS complex (median label width 216 to 364 ms per recording). In these recordings the macro-F1 against the cardiologist was 0.634, compared with 0.858 in the other 20 recordings. With the alternative 8–18 Hz detector band, the pooled LVAD-DB macro-F1 was 0.878 and the ranking of the four models was unchanged (Supplementary Table S6). The R-peak detector reached a sensitivity and PPV of 0.947 and 0.996 on LUDB, 1.000 and 0.999 on the biomedical engineer's recordings, and 0.992 and 0.932 on the cardiologist's segments, where most unmatched detections were double detections of the wide complexes (Supplementary Section S3). To quantify inter-rater variability, the biomedical engineer additionally annotated a stratified random subset of 299 of the cardiologist's beats (13 per patient, all 23 patients), blinded to the cardiologist's labels and to the model output. Twenty-one beats that the engineer marked as uninterpretable were excluded, leaving 278. On these beats the two experts agreed with a macro-F1 of 0.784 (95% CI 0.728 to 0.829, κ 0.711), whereas the model reached 0.836 (0.795 to 0.871) against the cardiologist and 0.884 (0.856 to 0.909) against the engineer. Outside the six wide-complex recordings the experts agreed with a macro-F1 of 0.834 (0.804 to 0.856) and the model with 0.866 and 0.887. Within them the cardiologist labelled a P wave in 9 of 58 beats and the engineer in all 58, so the experts' agreement fell to 0.554, while the model agreed with the engineer at 0.870. The spread of boundary timing between the experts (SD 20 to 33 ms per boundary) was similar to that between the model and the experts on LVAD-DB (18 to 38 ms). On these beats the model therefore agreed with each expert at least as closely as the experts agreed with each other.

The boundary timing statistics are conditional on a matched boundary (Fig. 2e, Supplementary Table S7). Of the expert boundaries, 95.5% were matched on LUDB and 91.7% on LVAD-DB, with per-boundary matched fractions of 90.3% to 98.5% and 81.2% to 99.0%, the lowest at the P onset. For matched boundaries, the bias pooled over all boundaries was $+0.6 \pm 19.1$ ms on LUDB and $+5.3 \pm 27.3$ ms on LVAD-DB. On LUDB the bias at the P and QRS boundaries and at the T offset ranged from -6.6 to $+0.5$ ms, with SDs of 10.4 to 22.5 ms. The QRS offset was the most precise boundary (-1.2 ± 10.4 ms), and the QRS onset was placed 6.6 ms early, in line with the anchor 15 samples before the detected R peak. On LVAD-DB the corresponding biases ranged from -3.8 to $+6.6$ ms. The T onset was the widest boundary on both datasets ($+12.8 \pm 25.6$ ms on LUDB, $+25.9 \pm 37.7$ ms on LVAD-DB), consistent with the shallow slope at the beginning of the T wave.

C. Clinical Intervals

Table II compares the clinical intervals derived from the predicted delineation with those derived from the expert annotations. The interval availability was 78.0% to 97.1% of the beats on LUDB and 65.5% to 81.5% on LVAD-DB, lowest for the P-wave intervals. Per record, the model's median QRS duration exceeded the expert's by $+12.7 \pm 11.8$ ms and its median QT by $+11.3 \pm 21.5$ ms on LUDB (196 records each), a

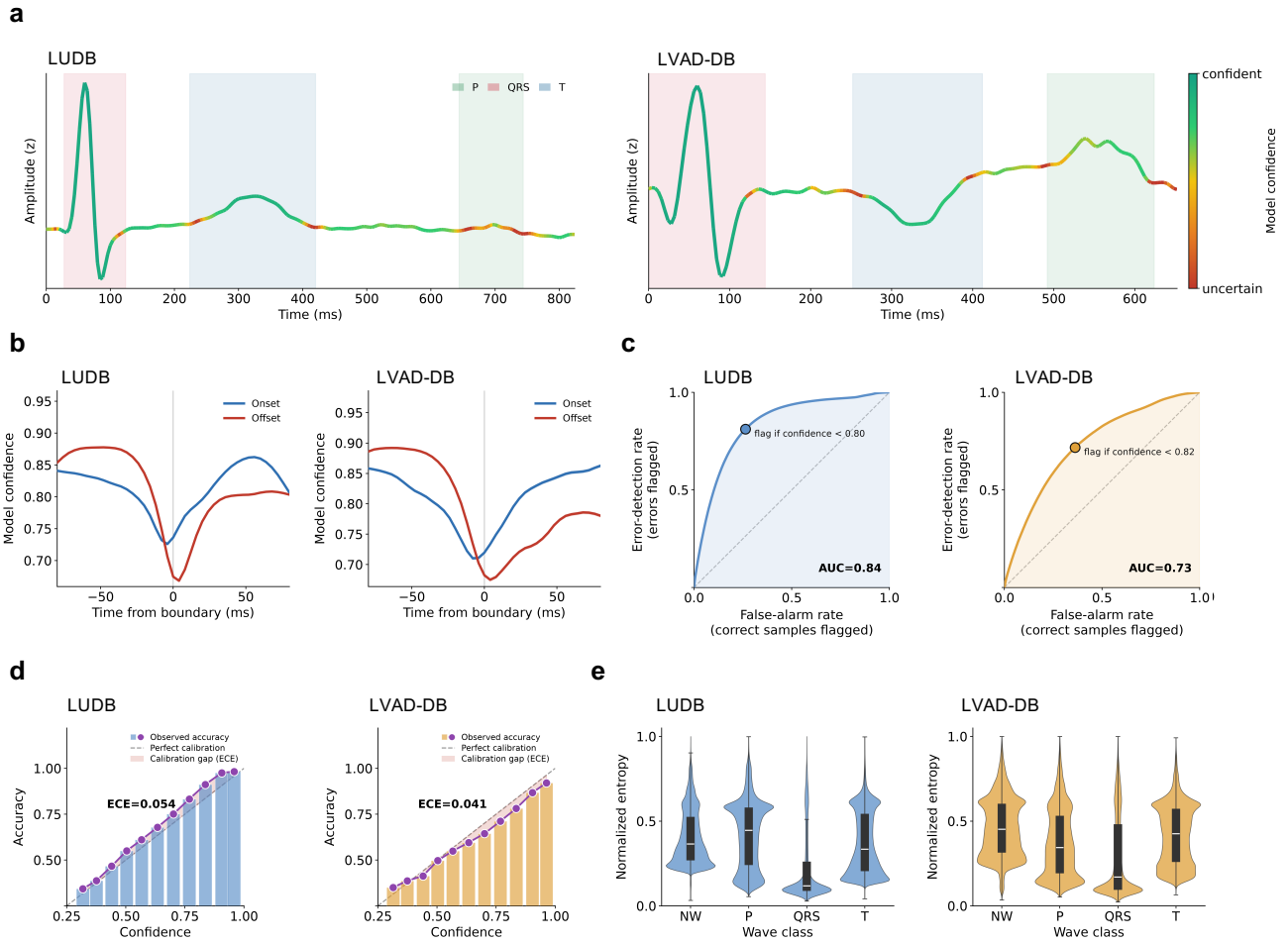


Fig. 3. Model trustworthiness on LUDB (left sub-plot of each panel) and LVAD-DB (right, pooled). **(a)** Per-sample confidence (maximum softmax probability, teal for confident and red for uncertain samples) along a representative beat. Shaded bands mark the ground-truth P, QRS, and T. **(b)** Confidence against time from onsets and offsets, averaged over all boundaries. Within ± 1 sample of a boundary the confidence was 0.718 (LUDB) and 0.738 (LVAD-DB), against wave-interior means of 0.838 and 0.824. **(c)** Confidence as an error detector. ROC curves for flagging misclassified samples, with Youden points (thresholds 0.80 and 0.82) and AUC of 0.836 and 0.732. **(d)** Reliability diagrams of the raw confidence against the original labels, with accuracy per confidence bin (bars), mean confidence (dots), and perfect calibration (dashed). The floor at 0.25 is the four-class softmax minimum, and ECE $= \mathbf{0.054}$ and $\mathbf{0.041}$. **(e)** Per-class normalized predictive entropy.

TABLE II

AGREEMENT OF MODEL-DERIVED AND EXPERT-DERIVED CLINICAL INTERVALS PER BEAT (MODEL − EXPERT, MS). BIAS ± SD, MEAN ABSOLUTE ERROR (MAE), AND AVAILABILITY (AVAIL.), THE FRACTION OF BEATS WITH THE INTERVAL MEASURABLE IN BOTH SOURCES. LVAD-DB POOLED ACROSS RATERS.

	LUDB			LVAD-DB		
Interval	**Bias ± SD**	**MAE**	**Avail.**	**Bias ± SD**	**MAE**	**Avail.**
P duration	-1.0 ± 34.0	24.4	78.0%	-19.2 ± 38.1	29.9	65.5%
PR interval	-14.6 ± 28.3	22.4	78.0%	-20.3 ± 38.2	28.9	65.5%
QRS duration	$+12.8 \pm 28.2$	20.9	97.1%	-3.4 ± 50.2	32.1	81.5%
QT interval	$+9.8 \pm 43.8$	28.4	95.2%	$+1.9 \pm 61.8$	39.0	72.6%
QTc (Bazett)	$+10.7 \pm 47.6$	30.8	94.4%	$+2.4 \pm 73.3$	46.9	71.6%
QTc (Fridericia)	$+10.4 \pm 46.1$	30.0	94.4%	$+2.1 \pm 69.2$	44.0	71.6%

systematic offset consistent with the early QRS onset reported above. Per beat, the MAE on LUDB ranged from 20.9 ms for QRS duration to 30.8 ms for QTcB, and on LVAD-DB from 28.9 ms for the PR interval to 46.9 ms for QTcB. These per-beat errors, in particular the LVAD-DB QTcB MAE of 46.9 ms with an SD of 73.3 ms, exceed the differences that are clinically relevant for QTc, so the model-derived intervals are not interchangeable with expert measurements. On LVAD-DB the per-record QRS duration was underestimated against the cardiologist (-26.5 ± 72.1 ms) but not against the biomedical engineer ($+8.0 \pm 20.2$ ms), reflecting the wide complexes described above. In 170 to 172 LUDB records in sinus rhythm, model and expert agreed on whether the record median lay within the normative range in 80.6% to 97.1% of records, with κ of 0.35 for QRS duration and 0.58 for QTcB (Supplementary Tables S8 and S9).

D. Calibration, Uncertainty, and Failure Modes

The raw softmax confidence had an ECE of 0.054 (95% CI 0.050–0.058) on LUDB and 0.041 (0.018–0.070) on LVAD-DB against the original labels (Table I, Fig. 3d). The temperature fitted on the QTDB validation records ($T = 0.71$) sharp-

TABLE III
COMPARISON WITH PUBLISHED ARCHITECTURES TRAINED WITH THE SAME WINDOWING, AUGMENTATION, AND PROCEDURE (MACRO-F1, LVAD-DB POOLED). IN PARENTHESES, Δ = OURS − BASELINE. *HOLM-ADJUSTED $p < 0.05$, PAIRED PATIENT-LEVEL BOOTSTRAP.

		Tolerance-aware		Strict	
Model (Ref.)	Params (k)	LUDB	LVAD-DB	LUDB	LVAD-DB
Bi-mLSTM (this study)	**2.5**	**0.919**	**0.875**	**0.850**	**0.761**
Peimankar et al. [5]	1,419	0.911 (+0.008*)	0.849 (+0.025*)	0.867 (−0.016*)	0.738 (+0.023)
Jiménez-Pérez et al. [7]	1,296	0.896 (+0.023*)	0.839 (+0.035*)	0.848 (+0.002)	0.724 (+0.037)
Liu et al. [11]	301	0.915 (+0.004*)	0.866 (+0.008)	0.869 (−0.019*)	0.758 (+0.002)

ened the confidence. It reduced the ECE on LUDB to 0.013 but increased it on LVAD-DB to 0.107 (Supplementary Table S10), so this post-hoc correction did not transfer under the domain shift to LVAD-DB. Low confidence detected the model's own misclassifications with an AUC of 0.836 (LUDB) and 0.732 (LVAD-DB) (Fig. 3c). At the Youden-optimal thresholds it flagged 81% and 72% of misclassified samples at the cost of flagging 26% and 36% of correct ones. Misclassified samples carried higher normalized predictive entropy than correct ones (0.57 vs. 0.35 on LUDB, 0.53 vs. 0.36 on LVAD-DB). Within a beat, confidence was highest on the QRS complex (mean 0.906 on LUDB, 0.864 on LVAD-DB) and dropped at wave boundaries, more at offsets than onsets on LUDB (0.672 vs. 0.763), consistent with the gradual return of the T wave to the isoelectric line (Fig. 3a,b).

The dominant failure mode on LVAD-DB was not a missed QRS complex but confusion between the broad, low-amplitude T wave and the adjacent isoelectric segment, together with premature or delayed T-wave boundaries. Uncertainty tracked these errors. Per-class entropy was lowest on the QRS complex on both datasets and, on LVAD-DB, highest on the NW and T-wave classes (Fig. 3e), so it could gate low-confidence beats for human review.

E. Comparison with Published Architectures

Trained with the same windowing and procedure, the 2,522-parameter model exceeded all three baselines in tolerance-aware macro-F1 on LUDB and matched the strongest of them on LVAD-DB (Table III, Fig. 2g, Supplementary Tables S11 and S12). Its margins over Peimankar et al. and Jiménez-Pérez et al. were significant on both datasets ($+0.008$ and $+0.023$ on LUDB, $+0.025$ and $+0.035$ on LVAD-DB). Over Liu et al., whose model has 119 times more parameters, the margin was small but significant on LUDB ($+0.004$, 95% CI $+0.002$ to $+0.006$, Holm-adjusted $p = 0.006$) and not significant on LVAD-DB ($+0.008$, 95% CI -0.001 to $+0.018$, adjusted $p = 0.067$). Our model was more accurate on the NW, QRS, and T-wave classes, whereas Peimankar et al. and Liu et al. attained higher P-wave F1 on LUDB (0.882 and 0.870 vs. 0.858), and the P-wave differences on LVAD-DB were not significant (Fig. 2d). Under strict scoring these two baselines also exceeded our model on LUDB (-0.016 and -0.019 macro-F1), mainly at the P wave, and no strict macro-F1 difference on LVAD-DB was significant after correction. In their native sliding-window pipelines the baselines reached 0.889–0.903 on LUDB and 0.721–0.723 on LVAD-DB (Supplementary Table S13). Because these pipelines also differ in checkpoint, training data, and preprocessing, the comparison is descriptive.

The comparison also concerns the hardware budget. Our model runs within the measured on-chip budget of the Arduino Portenta H7 without quantization (Fig. 4d, Supplementary Table S14). In 32-bit floating point the weights of the baselines alone (1.2–5.7 MB) exceed the 1 MB on-chip SRAM, and two of the three still do after 8-bit quantization (1.30 and 1.42 MB), so on this device they would require external memory or, for the model of Liu et al., 8-bit quantization.

F. On-Device Performance

On the Arduino Portenta H7 (Cortex-M7), the full per-beat pipeline ran in 128.6 ± 1.5 ms (preprocessing 1.5 ms, network inference 127.1 ± 1.5 ms, 469 beats), using 22 kB of the 1 MB on-chip SRAM (2.2%) and 780 kB of the 2 MB flash (38%). The pilot firmware ran an earlier set of weights of the identical architecture, so these latency and memory figures apply unchanged, and the on-device accuracy reported below is that of those weights (firmware verification in Supplementary Table S15). The 128.6 ms is processing time. Because each beat is classified once the following R peak is confirmed, its labels became available about one RR interval after its onset, plus the R-peak confirmation and the processing time, which amounts to about 0.9 s at the pilot median RR interval of 0.77 s. Two of the 469 pilot windows exceeded 300 samples and were truncated. Replaying the pilot through a bit-exact reference of the firmware classifier with per-window instead of buffer-level normalization changed macro-F1 by -0.004 for the deployed weights and by -0.009 for the final weights (Supplementary Table S15).

Worn on the chest (Fig. 4a–c) during a continuous 361 s rest–walk–rest recording, the device delineated 469 beats in real time (Fig. 4e). The logged output reached a macro-F1 of 0.93 and a κ of 0.89 pooled across both raters (0.93 and 0.90 against the biomedical engineer, 0.92 and 0.89 against the cardiologist), with per-class F1 of 0.95 (P), 0.89 (QRS), and 0.95 (T). Emulated through the firmware path on the same windows, the final weights reached a macro-F1 of 0.949 and a κ of 0.925, with per-class F1 of 0.94 (NW), 0.94 (P), 0.95 (QRS), and 0.97 (T) (Fig. 4g–i; Supplementary Table S15). Macro-F1 was 0.95 during the first seated segment, 0.93 during walking, and 0.91 during the second seated segment, and confidence decreased only slightly during walking (Fig. 4f).

IV. DISCUSSION

A. Transparent, Parameter-Efficient Delineation at the Edge

Deep delineators such as the U-Net-like model of Jiménez-Pérez et al. [7] and the RNNs of Peimankar et al. [5] and Nurmaini et al. [6] achieve >99% QRS F1 on LUDB and have surpassed wavelet- and rule-based methods [1]–[3], [8]–[15], but they remain too large for real-time use on microcontrollers. Compact edge models have so far targeted beat classification. ANNet [27] reached 97% accuracy on MIT-BIH and, after 8-bit quantization and Cortex-M4 deployment, reduced power

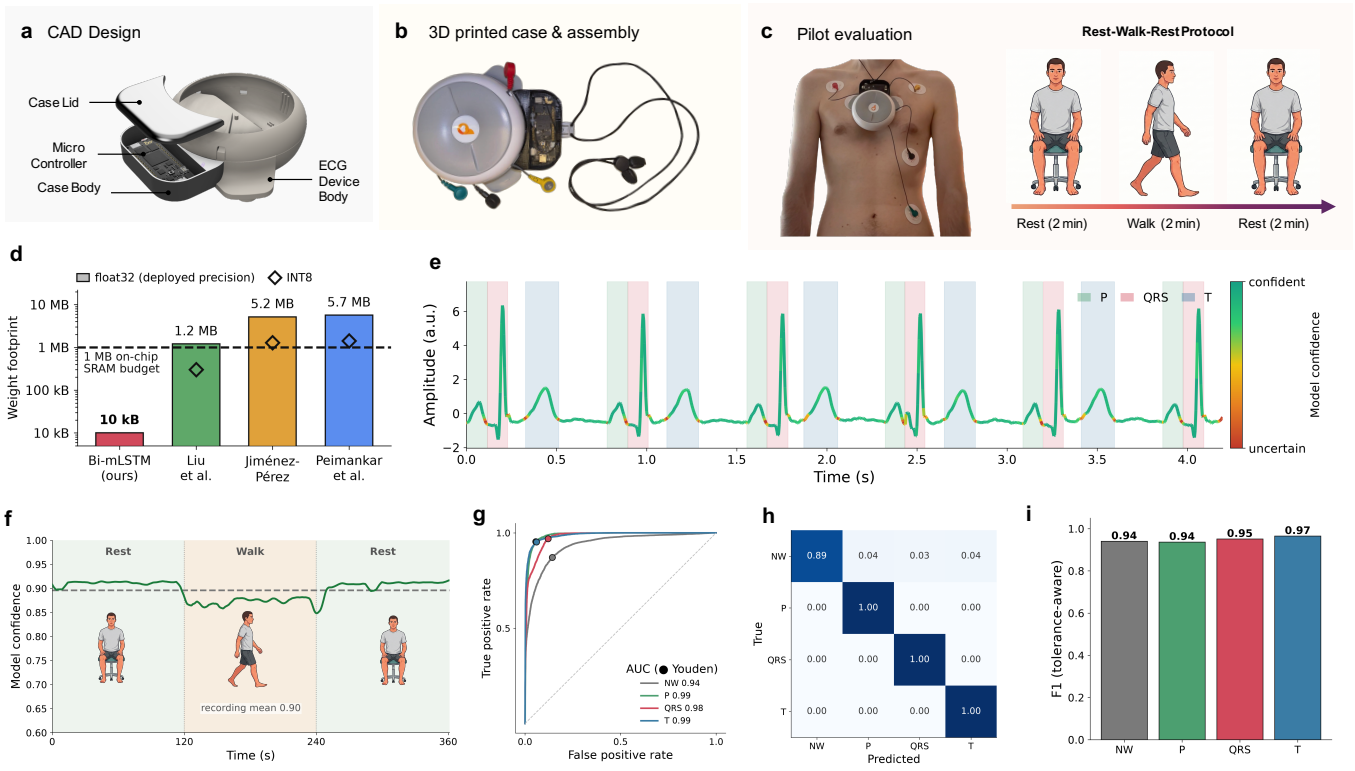


Fig. 4. End-to-end edge deployment in the pilot recording (N=1). **(a)** CAD design (exploded view) of the 3-D-printed enclosure for the Arduino Portenta H7 (Cortex-M7). **(b)** Printed box and assembled device. **(c)** Device worn on the chest (lead D2) during a protocol of 2 min seated rest, 2 min walking, and 2 min seated rest. **(d)** Model weight footprint (log scale) in 32-bit floating point (bars) against the 1 MB on-chip SRAM budget (dashed), 10 kB for the selected model and 1.2–5.7 MB for the baselines. Hollow diamonds mark the baseline INT8 sizes (0.30–1.42 MB). **(e,f)** Device run with the deployed weights. **(e)** Delineation of the live 250 Hz stream (469 beats, 361 s) as logged by the device, colored by confidence (teal for confident, red for uncertain samples) over the predicted class. **(f)** Confidence (5 s edge-corrected rolling mean) across protocol phases. The confidence in (e,f) is the softmax output of the firmware-path replay of the deployed weights, whose labels are identical to the logged output. **(g–i)** Final model of the main text emulated through the same firmware path on the device's beat windows and input (Supplementary Table S15), an emulation rather than an on-device measurement, scored over the five annotated 10-s segments with both raters pooled. **(g)** Per-class ROC curves with Youden points. **(h)** Row-normalized tolerance-aware confusion matrix. **(i)** Per-class tolerance-aware F1 (macro-F1 0.949, κ 0.925). Panels (g–i) are single-subject point estimates.

by transmitting only pathological beats. The classifiers of Rahman and Morshed [18] and Sarhaddi et al. [19] likewise predict one label per beat or record, so their latencies are not directly comparable to the dense per-sample delineation of a 300-sample window produced by Bi-mLSTM. These models are task-specific black boxes, and a new diagnostic question requires a new model. Delineation instead exposes the physiology. The P-wave duration, PR interval, QRS duration, and QTc derived from our model are the measures that clinicians read, trust, and can verify against the waveform. Their per-record agreement with the expert-derived values (Table II) supports trend monitoring, whereas the per-beat errors still preclude their use in place of expert measurement. The same intervals can feed simple, interpretable models for many diagnostic tasks, which makes on-device delineation a reusable and transparent front end for wearable monitoring.

Within delineation, a 2,522-parameter network matched or exceeded architectures more than 100 times larger under identical windowing and training in tolerance-aware scoring, and it generalized to pathological LVAD morphology and to two independent annotators. Its relative weakness is the P wave, where two baselines are more accurate on LUDB (Fig. 2d), which also gives them the higher strict macro-F1 on LUDB.

Real-time QRS detection on edge devices was demonstrated by Gradl et al. [28], and consumer wearable devices now detect atrial fibrillation at scale [29], [30]. As AI-ECG extends to screening for structural and ischemic disease [31]–[37], the need for efficient models that run where the signal is acquired grows. Our pilot shows that a delineator with fewer than 5,000 parameters fits comfortably within a Cortex-M7 envelope while streaming from a commercial portable monitor.

B. Beat Anchoring as a Design Choice

Placing each window at a detected R peak is a deliberate design choice. R-peak detection is a mature and robust task with negligible computational cost, and it already runs on wearable ECG devices [28]. Anchoring gives the network a consistent temporal reference and the context of one complete cardiac cycle, so the network does not have to learn invariance to arbitrary window positions. The controlled comparison of the same architecture on anchored and unanchored windows supports this argument. With identical weights, the macro-F1 of our model fell from 0.919 to 0.834 on LUDB and

from 0.875 to 0.654 on LVAD-DB when the beat-anchored windows were replaced by unanchored ones, and without anchor augmentation it fell to 0.707 and 0.583 (Supplementary Table S3). This comparison does not include a model trained on unanchored windows, so it quantifies the dependence of the trained model on anchoring rather than the full benefit of anchoring as a design choice. The native-pipeline baselines differ in more than window placement and therefore do not isolate this effect. The price of anchoring is a dependency on the detector. Its misses, false detections, and anchor jitter are included in every reported number, its sensitivity was at least 0.947 on every external dataset, and augmentation with displaced anchors makes the model tolerant of the remaining jitter.

C. Uncertainty and Calibration

The model's confidence localizes to physiological structure. It is high through the QRS complex and lower at the P- and T-wave boundaries, where delineation is intrinsically hardest, and the raw confidence had an ECE of at most 0.054 on both external sets. Vranken et al. found, across 500,000 ECGs, that cardiologist disagreement often co-occurred with model uncertainty [38]. Similarly, LVAD ECGs lacking clear P waves raised our model's uncertainty, and low confidence flagged its own errors. Calibration can deteriorate under class imbalance even when discrimination improves [39], and post-hoc recalibration can fail under domain shift. A temperature fitted on QTDB improved the calibration on LUDB but degraded it on LVAD-DB, so a target-domain calibration set would be needed before recalibrated probabilities are used in a new population. Interpretable clinical features, uncertainty quantification, and calibration together align with best practice in clinical AI development [40].

D. Limitations and Future Directions

The model was trained on automatic QTDB boundaries rather than manual annotations, and its external scores are therefore a test of agreement between automatically supervised predictions and expert labels. Offline, the R-peak detector searched only the annotated stretches, so its PPV does not include false detections in unannotated signal. The early QRS onset produced systematic positive biases in the LUDB QRS duration and QT, and per-beat interval errors on LVAD-DB remain too large for diagnostic interval measurement. The on-device pilot is a single, non-pathological recording from one healthy adult, acquired with a separately trained earlier version of the same architecture. On-device robustness to pathological morphology and to a broader range of motion and noise remains to be established in the planned prospective study with larger cohorts. Inter-rater agreement was measured on a stratified subset of 278 beats from one pair of experts, and the remaining LVAD-DB beats were annotated by a single expert each, so the per-rater results also reflect differences between the annotated beats. The model is single-lead. Multi-lead variants could capture inter-lead dynamics without drastically increasing model size, few-shot personalization could improve robustness to physiological or electrode drift, and transmitting only anomalous beats, as in ANNet [27], could further reduce energy demands.

V. Conclusion

A recurrent network with only 2,522 parameters delineates the P wave, QRS complex, and T wave of the ECG with an accuracy comparable to that of published models more than 100 times larger. It exceeded the strongest baseline on LUDB and matched it on LVAD-DB in tolerance-aware scoring. It reports how confident it is in every sample and yields the clinical intervals that clinicians already use, with per-record agreement suited to trend monitoring. The network runs in real time within the on-chip memory of a small microcontroller attached to a portable ECG monitor. Accurate and transparent ECG analysis is therefore possible directly on wearable and other resource-constrained devices.

Data and Code Availability

Source code for preprocessing, training, evaluation, and the real-time deployment is available under the GPLv2 license at `https://github.com/CellularSyntax/BixECG`, with a frozen archive at `https://doi.org/10.5281/zenodo.23196330`; the trained model weights are available at `https://huggingface.co/maxhaberbusch/BixECG`. QTDB [24] and LUDB [25], [26] are openly available via PhysioNet. The LVAD-DB and D-Heart pilot recordings cannot be shared publicly due to data-protection regulations but are available from the corresponding author on reasonable request.

[1] J. M. Bote, J. Recas, F. Rincn, D. Atienza, and R. Hermida, "A Modular Low-Complexity ECG Delineation Algorithm for Real-Time Embedded Systems," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 429–441, 2018.

[2] P. Laguna, R. Jané, and P. Caminal, "Automatic Detection of Wave Boundaries in Multilead ECG Signals: Validation with the CSE Database," *Comput. Biomed. Res.*, vol. 27, no. 1, pp. 45–60, 1994.

[3] J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ECG delineator: evaluation on standard databases," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 570–581, 2004.

[4] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. E215–20, 2000.

[5] A. Peimankar and S. Puthusserypady, "DENS-ECG: A deep learning approach for ECG signal delineation," *Expert Syst. Appl.*, vol. 165, p. 113911, 2021.

[6] S. Nurmaini *et al.*, "Robust electrocardiogram delineation model for automatic morphological abnormality interpretation," *Sci. Rep.*, vol. 13, no. 1, p. 13736, 2023.

[7] G. Jimenez-Perez, J. Acosta, A. Alcaine, and O. Camara, "Generalising electrocardiogram detection and delineation: training convolutional neural networks with synthetic data augmentation," *Front. Cardiovasc. Med.*, vol. 11, p. 1341786, 2024.

[8] P. Sodmann, M. Vollmer, N. Nath, and L. Kaderali, "A convolutional neural network for ECG annotation as the basis for classification of cardiac rhythms," *Physiol. Meas.*, vol. 39, no. 10, p. 104005, 2018.

[9] A. Darmawahyuni *et al.*, "Improved delineation model of a standard 12-lead electrocardiogram based on a deep learning algorithm," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 139, 2023.

[10] X. Liang *et al.*, "ECG_SegNet: An ECG delineation model based on the encoder-decoder structure," *Comput. Biol. Med.*, vol. 145, p. 105445, 2022.

[11] J. Liu *et al.*, "A novel P-QRS-T wave localization method in ECG signals based on hybrid neural networks," *Comput. Biol. Med.*, vol. 150, p. 106110, 2022.

[12] X. Li, W. Cai, B. Xu, Y. Jiang, M. Qi, and M. Wang, "SEResUTer: a deep learning approach for accurate ECG signal delineation and atrial fibrillation detection," *Physiol. Meas.*, vol. 44, no. 12, p. 125005, 2023.
[13] D. Wang *et al.*, "Inter-patient ECG characteristic wave detection based on convolutional neural network combined with transformer," *Biomed. Signal Process. Control*, vol. 81, p. 104436, 2023.
[14] G. Jimenez-Perez, A. Alcaine, and O. Camara, "U-Net Architecture for the Automatic Detection and Delineation of the Electrocardiogram," in *2019 Comput. Cardiol. (CinC)*, 2019, pp. 1–4.
[15] ——, "Delineation of the electrocardiogram with a mixed-quality-annotations dataset using convolutional neural networks," *Sci. Rep.*, vol. 11, no. 1, p. 863, 2021.
[16] S. Nurmaini *et al.*, "Beat-to-Beat Electrocardiogram Waveform Classification Based on a Stacked Convolutional and Bidirectional Long Short-Term Memory," *IEEE Access*, vol. 9, pp. 92 600–92 613, 2021.
[17] T. Meuser *et al.*, "Revisiting Edge AI: Opportunities and Challenges," *IEEE Internet Comput.*, vol. 28, no. 4, pp. 49–59, 2024.
[18] M. Rahman and B. I. Morshed, "Resource-Constrained On-Chip AI Classifier for Beat-by-Beat Real-Time Arrhythmia Detection with an ECG Wearable System," *Electronics*, vol. 14, no. 13, p. 2654, 2025.
[19] F. Sarhaddi, Y. Wang, M. Khan, N. H. Motlagh, A. Zuniga, H. Flores, and P. Nurmi, "Spiking Recurrent ECG Classification for Resource-Constrained Wearable Devices," in *Proc. IEEE-EMBS Int. Conf. Body Sensor Netw. (BSN)*, Jul. 2026.
[20] M. Beck *et al.*, "xLSTM: Extended Long Short-Term Memory," arXiv preprint, 2024.
[21] L. Kang, X. Fu, J. Vazquez-Corral, E. Valveny, and D. Karatzas, "xLSTM-ECG: Multi-label ECG Classification via Feature Fusion with xLSTM," arXiv preprint, 2025.
[22] S. Datta, R. K. Thondithara, and S. Kumar, "Synthesizing 12-Lead ECG from Reduced Lead Sets using Deep Neural Networks for Remote Healthcare Applications," *IEEE Sens. Lett.*, vol. PP, no. 99, pp. 1–4, 2025.
[23] R. Lunelli, A. Nicolson, S. M. Pröll, S. J. Reinstadler, A. Bauer, and C. Dlaska, "BenchECG and xECG: A Benchmark and Baseline for ECG Foundation Models," *npj Digit. Med.*, 2026.
[24] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in *Comput. Cardiol. 1997*, 1997, pp. 673–676.
[25] A. I. Kalyakulina *et al.*, "LUDB: A New Open-Access Validation Tool for Electrocardiogram Delineation Algorithms," *IEEE Access*, vol. 8, pp. 186 181–186 190, 2020.
[26] A. Kalyakulina, I. Yusipov, V. Moskalenko, A. Nikolskiy, K. Kosonogov, N. Zolotykh, and M. Ivanchenko, "Lobachevsky University Electrocardiography Database (version 1.0.1)," PhysioNet, 2021.
[27] G. Sivapalan, K. K. Nundy, S. Dev, B. Cardiff, and D. John, "ANNet: A Lightweight Neural Network for ECG Anomaly Detection in IoT Edge Sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 1, pp. 24–35, 2022.
[28] S. Gradl, P. Kugler, C. Lohmuller, and B. Eskofier, "Real-time ECG monitoring and arrhythmia detection using Android-based mobile devices," in *2012 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2012, 2012, pp. 2452–2455.
[29] M. V. Perez *et al.*, "Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation," *N. Engl. J. Med.*, vol. 381, no. 20, pp. 1909–1917, 2019.
[30] Y. Guo, H. Wang, H. Zhang, T. Liu, Z. Liang, Y. Xia, L. Yan, Y. Xing *et al.*, "Mobile Photoplethysmographic Technology to Detect Atrial Fibrillation," *J. Am. Coll. Cardiol.*, vol. 74, no. 19, pp. 2365–2375, 2019.
[31] W.-Y. Ko *et al.*, "Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram," *J. Am. Coll. Cardiol.*, vol. 75, no. 7, pp. 722–733, 2020.
[32] M. Cohen-Shelly *et al.*, "Electrocardiogram screening for aortic valve stenosis using artificial intelligence," *Eur. Heart J.*, vol. 42, no. 30, pp. 2885–2896, 2021.
[33] Z. I. Attia *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," *Nat. Med.*, vol. 25, no. 1, pp. 70–74, 2019.
[34] R. Alam, A. Aguirre, and C. M. Stultz, "Detecting QT prolongation from a single-lead ECG with deep learning," *PLOS Digit. Health*, vol. 3, no. 6, p. e0000539, 2024.
[35] R. Herman *et al.*, "Validation of an automated artificial intelligence system for 12-lead ECG interpretation," *J. Electrocardiol.*, vol. 82, pp. 147–154, 2024.
[36] ——, "International evaluation of an artificial intelligence–powered electrocardiogram model detecting acute coronary occlusion myocardial infarction," *Eur. Heart J. - Digit. Health*, vol. 5, no. 2, pp. 123–133, 2023.
[37] R. Khera *et al.*, "Transforming Cardiovascular Care With Artificial Intelligence: From Discovery to Practice JACC State-of-the-Art Review," *J. Am. Coll. Cardiol.*, vol. 84, no. 1, pp. 97–114, 2024.
[38] J. F. Vranken *et al.*, "Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms," *Eur. Heart J. - Digit. Health*, vol. 2, no. 3, pp. 401–415, 2021.
[39] G. Baj *et al.*, "Comparison of discrimination and calibration performance of ECG-based machine learning models for prediction of new-onset atrial fibrillation," *BMC Med. Res. Methodol.*, vol. 23, no. 1, p. 169, 2023.
[40] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019.