Supplementary Material

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

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.

S1. Hyperparameter search and architecture selection

The compact architecture was chosen from a 60-configuration multi-objective search over the xLSTM design space, jointly maximizing QTDB validation macro-F1 while minimizing parameter count. The search space is given in Table S1, representative configurations spanning the size range in Table S2, and the full search in Fig. S1. To keep the search inexpensive, every configuration was trained and evaluated on windows placed at the annotated QRS onsets of each dataset. The selected architecture was then trained on detector-placed windows (main-text Methods), so the search-time scores in this section differ from the final results. The architecture and the final checkpoint were selected on QTDB alone and frozen before any external evaluation. The external scores reported in this section were computed afterwards and did not inform the selection.

Table S1: Hyperparameter search space for the mLSTM architecture. The optimization maximized validation macro-F1 while minimizing model size. Fixed training parameters (optimizer, learning-rate schedule, loss function) are described in the main-text Methods.

Hyperparameter	**Search range / values**	**Description**
Number of blocks	2, 3, 4	Total stacked recurrent blocks
sLSTM layer positions	All k-combinations of block indices ($k = 1$–3)	Positions of simple LSTM layers among blocks
Number of heads	1, 2, 4	Attention-like heads per mLSTM block
Embedding size	8 to 100 (step 4)	mLSTM hidden-layer size
1D convolution kernel size	3, 7, 11, 21, 31, 41, 51, 61, 71	Temporal receptive field of CNN front-end
Sequence duration (s)	0.5, 0.75, 1.0, 1.2, 1.5	Window duration for beat-aligned input
Dropout rate	0.10 to 0.50 (step 0.05)	Applied after recurrent layers

S1.1. Optimization procedure

Multi-objective Bayesian optimization was performed with Optuna, jointly maximizing QTDB validation macro-F1 and minimizing parameter count. The search space (Table S1) included the

Table S2: Representative mLSTM configurations of the search and their cross-domain performance on QTDB (training) and the two external datasets. Model 1 is the selected architecture, which offers the best trade-off between size and performance. All values are search-time scores on windows placed at the annotated QRS onsets (tolerance-aware macro-F1, LVAD-DB pooled across both raters, QTDB column the validation macro-F1) and serve the comparison across configurations only. The selected architecture trained on detector-placed windows reaches 0.919 (LUDB) and 0.875 (LVAD-DB), main-text Table I.

Model	Params (k)	Train F1 (QTDB)	Test F1 (LUDB)	Test F1 (LVAD-DB)	Notes
1	2,522	0.88	0.94	0.91	Final selected model (QTDB Pareto knee)
2	1,140	0.81	0.82	0.71	Lightweight variant (smallest)
3	42,980	0.91	0.92	0.90	Overparameterized (largest)
4	5,478	0.88	0.89	0.85	Intermediate configuration

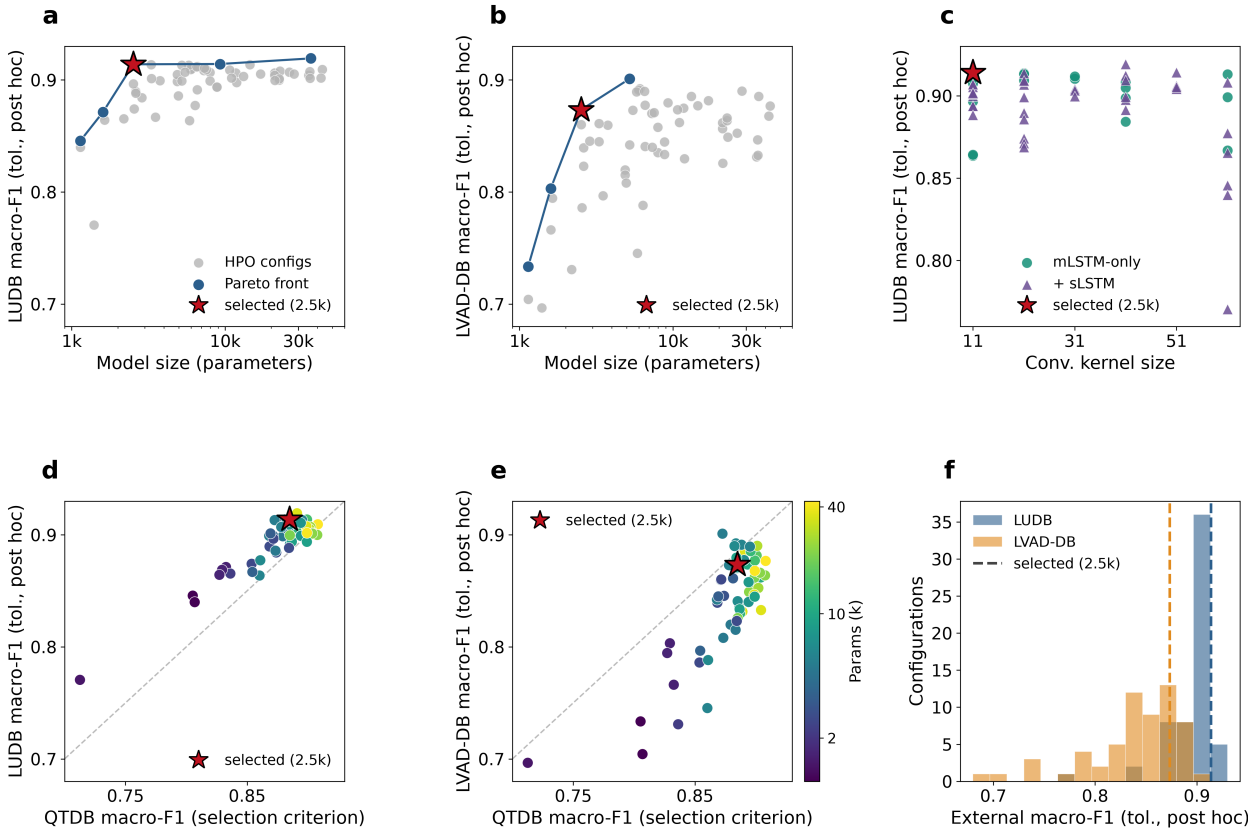


Figure S1: Hyperparameter search across 60 xLSTM configurations, each trained on QTDB. Selection used only the QTDB macro-F1 and the parameter count. The LUDB and LVAD-DB scores are post-selection analyses that played no part in the selection (tolerance-aware macro-F1, search-time scores on windows placed at the annotated QRS onsets). The selected 2,522-parameter configuration is marked by one red star per panel, and by dashed lines in (f). **(a,b)** Accuracy against model size on LUDB and LVAD-DB. The selected configuration lies on the accuracy–size Pareto front (blue) of both external sets. **(c)** LUDB macro-F1 against convolution kernel size, separated by whether the stack includes an sLSTM block. **(d,e)** External macro-F1 against QTDB macro-F1, colored by parameter count. External accuracy increases with QTDB accuracy. **(f)** Distribution of external macro-F1 across the 60 configurations. The selected configuration ranks 3rd on LUDB and 16th on LVAD-DB.

number of blocks (2–4), the relative placement of sLSTM layers within the block stack, the number of attention heads (1, 2, or 4), the convolution kernel size (3–71), the embedding dimension (8–100), the sequence duration (0.5–1.5 s), and the dropout rate (0.1–0.5). Each candidate configuration was trained from scratch on an 80/20 record-level QTDB split and evaluated by validation macro-F1. Training runs were pruned early with Optuna's median pruning strategy if no performance gains were observed, trials were run with a fixed seed to ensure reproducibility, and the batch size was adjusted dynamically to fit GPU memory constraints. A total of 60 trials were conducted. Model selection was performed entirely on the training domain. Configurations were ranked on the trade-off between QTDB validation macro-F1 and parameter count, the accuracy–size Pareto front was identified, and the model at its knee was selected. This is the smallest architecture that captures essentially all achievable training accuracy, beyond which further parameters yield negligible gains. The external datasets (LUDB and LVAD-DB) were not used for selection. Their scores below are post-selection analyses.

The 60 configurations spanned 1,140 to 42,980 parameters (median 6,576, interquartile range 3,457 to 15,137). Their QTDB validation macro-F1 was tightly clustered (median 0.885, interquartile range 0.872–0.895, mean 0.877), reflecting a consistently high-performing search space rather than a few isolated good configurations. Evaluated on the external datasets after the selection, the configurations reached an average macro-F1 of 0.89 (95% CI 0.89–0.90) on LUDB and 0.84 (95% CI 0.83–0.85) on LVAD-DB, with corresponding macro-AUC of 0.97 (LUDB) and 0.92 (LVAD-DB). External macro-F1 correlated strongly with QTDB training macro-F1 (LUDB $r = 0.95$, $p < 0.001$, LVAD-DB $r = 0.87$, $p < 0.001$; Fig. S1d,e). At search time the selected 2,522-parameter model reached a QTDB validation macro-F1 of 0.88 (Table S2). The same architecture was trained twice in the search, with dropout 0.1 and 0.4. The dropout-0.1 trial (QTDB macro-F1 0.885) lies on the Pareto front and is the one marked in Fig. S1, and the dropout-0.4 trial reached 0.872.

S2. Effect of augmentation

Table S3 compares the selected architecture trained on detector-placed QTDB windows without and with augmentation. On detector-placed windows the two models are equivalent (macro-F1 0.914 vs. 0.919 on LUDB and 0.881 vs. 0.875 on LVAD-DB). Augmentation improves robustness to misplaced windows. On unanchored windows, whose position relative to the beat is arbitrary, it raises macro-F1 from 0.707 to 0.834 on LUDB and from 0.583 to 0.654 on LVAD-DB. For the same weights, the difference between the detector-placed and the unanchored columns quantifies how strongly the trained model depends on beat anchoring. The augmented model is the one reported throughout.

S3. Implementation details

Training data. The QTDB signals and labels were taken from a redistribution of all 105 two-lead records that matches PhysioNet QTDB 1.0.0 sample by sample. The labels of the first lead are the automatic pu0 boundaries and those of the second lead the automatic pu1 boundaries, and the manual q1c and q2c annotations were not used. Over the 210 lead blocks, the detector found 219,954 R peaks. Of the 219,744 intervals between consecutive detections, 16,714 (7.6%) exceeded 300 samples and were skipped, and 875 further windows without a labelled QRS sample were removed, which left 202,155 clean windows. Record sel33 (median RR interval 1.7 s) contributed no window. The three jittered copies added 201,437, 201,545, and 201,476 windows (slightly fewer than the clean copy, because some displaced intervals exceeded 300 samples or missed the QRS

Table S3: Selected architecture trained on detector-placed QTDB windows without and with augmentation (tolerance-aware scores). Accuracy on detector-placed windows is equivalent, and augmentation improves robustness to misplaced windows.

	Detector-placed windows[a]				Unanchored windows[b]	
	LUDB		LVAD-DB[c]		LUDB	LVAD-DB[c]
Training	**Macro-F1**	**QRS F1**	**Macro-F1**	**QRS F1**	**Macro-F1**	**Macro-F1**
No augmentation	0.914	0.910	0.881	0.911	0.707	0.583
With augmentation (final model)	0.919	0.925	0.875	0.903	0.834	0.654

[a]Anchor at the detected R peak minus 15 samples, 5–15 Hz detector band on every dataset (main-text Methods). [b]Non-overlapping 300-sample (1.2 s) windows at a fixed stride over the whole annotated stream, without R-peak or onset anchoring, so that the window position relative to the beat is arbitrary. Same 0.5–40 Hz band-pass, per-window z-normalization, and class-wise tolerance (P 52, QRS 48, T 124 ms). [c]Both raters pooled. Both models share the same training procedure, and the augmented model adds anchor jitter ($\sigma = 40$ ms, clipped at ± 60 ms, three copies per beat), time stretching (± 10 %), baseline wander, and Gaussian noise.

complex), for a total of 806,613. During training, each anchor was the detected R peak minus the R-to-onset delay of its record and lead, estimated from the training labels as the median position of the steepest slope in the first 80 ms after the labelled QRS onsets. These delays ranged from 7 to 19 samples, and their median over the 166 lead blocks of the 83 training records, 15 samples, is the constant used for every external evaluation. The record-level split assigned 83 records (633,306 windows) to training and 21 records (173,307 windows) to validation. The European ST-T records sele0121/sele0122, sele0124/sele0126, and sele0129/sele0133 come from one subject each and are all in the training partition.

R-peak detector. The Pan–Tompkins-style detector comprises a second-order Butterworth band-pass (5–15 Hz on every dataset), derivative, squaring, 150-ms moving-window integration, and adaptive signal and noise thresholds with a 200-ms refractory period and a search-back for missed beats at 1.66 times the median RR interval. Each detection is refined to the largest absolute deflection within ± 50 ms. Offline, the detector was run separately on each annotated stretch, defined by splitting the annotated QRS onsets at gaps longer than 1.5 s and padding each stretch by 0.4 s on either side, and its thresholds were re-initialized on every stretch (153 stretches on LUDB, 5 on the engineer's and 392 on the cardiologist's LVAD-DB recordings). Detections outside the stretches are therefore impossible, and the detector PPV does not reflect false detections in unannotated signal. Windows started by a detection in the padding or by a false detection were scored with the reference labels, which carry NW outside the waves. On LUDB, 9 windows started in the padding and 67 windows contained no labelled QRS sample, and excluding them would change macro-F1 from 0.919 to 0.920. A missed beat produces an interval longer than 300 samples, which is excluded, so missed beats lower the number of scored windows and are penalized through the detector sensitivity. Windows are z-score normalized on a per-window basis after cutting. As a sensitivity analysis, LVAD-DB was also evaluated with an alternative 8–18 Hz detector band (Table S6).

Detector scoring. Detector performance was scored against a reference R peak placed at the largest deflection within each annotated QRS complex, with a 150-ms matching tolerance. The cardiologist annotated partially labelled 10-s excerpts, so only detections inside the annotated stretches were scored. Under this definition the detector reached a sensitivity and PPV of 0.947 and 0.996 on LUDB, 1.000 and 0.999 on the engineer's recordings, and 0.992 and 0.932 on the

cardiologist's segments. Of the 241 remaining unmatched detections on the cardiologist's segments, 168 occurred in the six recordings in which wide, biphasic ventricular complexes were labelled as QRS, where the detector responded to both deflections of a complex. Signal quality did not differ between excerpts with and without missed beats (median signal-to-noise ratio 32.5 and 32.0 dB).

Network. The architecture below was read from the trained checkpoint (47 tensors, 2,522 parameters, loaded with strict key matching). A recurrent reimplementation in NumPy reproduces the PyTorch logits to within 5×10^{-6}. A linear layer maps each input sample to an 8-dimensional embedding (16 parameters). Three identical mLSTM blocks (810 parameters each) follow, each with a pre-normalization residual structure, $x \leftarrow x + f(\mathrm{LN}(x))$. Inside a block, the normalized input is projected up to 32 channels and split into a 16-channel cell path and a 16-channel output gate z. The cell path passes through a causal depthwise convolution with kernel 11 and a SiLU activation. Queries and keys are block-diagonal projections (eight 2×2 blocks) of the convolved path, and values are the same projection of the unconvolved path. Scalar input and forget gates are linear functions of the concatenated queries, keys, and values. The cell updates a 16×16 matrix memory and a normalizer state with an exponential input gate, a sigmoid forget gate, and log-domain stabilization, and keys are scaled by $1/\sqrt{16}$. The cell output is normalized per head, a learnable channel-wise skip of the convolved path is added, and the result is gated by $\mathrm{SiLU}(z)$ and projected back down to 8 channels. A final layer normalization follows the last block. All layer normalizations have no bias and a gain parametrized as $1 + w$ ($\epsilon = 10^{-5}$). The same stack, with shared weights, is applied to the window and to its time reversal. The two normalized 8-dimensional outputs are concatenated, and dropout ($p = 0.4$, training only) and a linear layer (68 parameters) produce four logits per sample, followed by a softmax. There is no separate convolutional stem and no feedforward sublayer.

Training. Padded samples carry an ignore label and are excluded from the loss and from all metrics. Time-stretch augmentation used nearest-neighbour label resampling, baseline wander had a random phase and an amplitude drawn uniformly up to 0.3 z-units, and the Gaussian-noise standard deviation was drawn uniformly from 0 to 0.15 z-units. The record-level 80/20 split used a fixed seed. Optimization used Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, learning rate 10^{-3}) with a two-epoch warm-up followed by cosine annealing to 10^{-5} over 12 epochs (equivalent to about 48 passes over the un-augmented beats), batch size 256, and early stopping (patience 8 epochs) on the validation tolerance-aware macro-F1, retaining the best checkpoint. Mixed precision was disabled to ensure reproducibility.

Firmware. The D-Heart monitor streams single-lead (D2) ECG at 160 Hz in 20-byte delta-encoded notification frames (24-bit ADC), which the board decodes to millivolts. The band-pass filter is applied forward and backward. The filtered signal is z-normalized once per 5-s analysis buffer, and each beat window is copied from the normalized buffer and zero-padded to 300 samples without further normalization. A beat is classified as soon as the following R peak has been detected, from a 1.2-s window buffer. Windows longer than 300 samples are truncated to their first 300 samples, and only the labels of the covered samples are emitted. Three concurrent RTOS contexts run BLE handling (at above-normal priority), the signal processing and inference pipeline, and host output and logging. Per-sample P, QRS, and T labels are stored in the board's QSPI flash memory or, optionally, streamed to a host over WiFi (TCP) or re-advertised over BLE. All components were tested with live and replayed ECG streams.

Software. Python (v3.10), PyTorch (v2.1.0), and CUDA (v11.8) on an NVIDIA GeForce RTX 3070 Ti GPU. The public repository also includes evaluation scripts, example inference outputs, and the Portenta H7 deployment scripts.

S4. Per-class diagnostic metrics

Table S4 reports the full per-class breakdown (PPV, NPV, sensitivity, specificity, one-vs-rest AUC, and F1) of the selected model on both external datasets under strict (per-sample, exact-match) and tolerance-aware evaluation. Strict metrics compare the argmax prediction with the expert label at every non-padded sample. Tolerance-aware metrics compare the tolerance-matched prediction (main-text Evaluation section) with the same expert labels, so that both share one denominator, the expert-labelled samples of the scored windows. All counts are pooled over samples before the per-class metrics are computed. One-vs-rest AUC uses the predicted class probabilities against the original labels and is therefore identical in both blocks. Both macro-F1 values are given in main-text Table I.

Table S4: Complete per-class diagnostic metrics of the selected 2,522-parameter Bi-mLSTM model on the two external datasets under *strict* (per-sample, exact-match) and *tolerance-aware* evaluation (LVAD-DB pooled across both raters). PPV, positive predictive value. NPV, negative predictive value. Sens., sensitivity. Spec., specificity. AUC, one-vs-rest area under the ROC curve, computed from class probabilities against the original labels and therefore identical under both regimes. F1, per-class F1. Macro rows average the four classes. Tolerance windows are ±52 ms (P), ±48 ms (QRS), and ±124 ms (T).

Class	PPV	NPV	Sens.	Spec.	AUC	F1
LUDB (35,558 windows, 199 subjects), strict						
NW	0.895	0.860	0.898	0.856	0.950	0.896
P wave	0.744	0.974	0.755	0.972	0.980	0.749
QRS complex	0.844	0.991	0.934	0.976	0.993	0.887
T wave	0.907	0.958	0.833	0.978	0.985	0.869
Macro	0.848	0.946	0.855	0.946	0.977	**0.850**
LUDB (35,558 windows, 199 subjects), tolerance-aware						
NW	0.985	0.875	0.898	0.981	0.950	0.939
P wave	0.792	0.993	0.936	0.974	0.980	0.858
QRS complex	0.864	0.999	0.995	0.978	0.993	0.925
T wave	0.922	0.997	0.989	0.979	0.985	0.954
Macro	0.891	0.966	0.954	0.978	0.977	**0.919**
LVAD-DB (both raters pooled, 5,205 windows), strict						
NW	0.689	0.859	0.758	0.812	0.858	0.722
P wave	0.715	0.949	0.698	0.953	0.944	0.706
QRS complex	0.848	0.934	0.775	0.958	0.936	0.810
T wave	0.823	0.923	0.787	0.938	0.954	0.805
Macro	0.769	0.916	0.754	0.915	0.923	**0.761**
LVAD-DB (both raters pooled, 5,205 windows), tolerance-aware						
NW	0.925	0.879	0.758	0.966	0.858	0.833
P wave	0.800	0.980	0.886	0.962	0.944	0.841
QRS complex	0.887	0.975	0.919	0.965	0.936	0.903
T wave	0.862	0.996	0.990	0.942	0.954	0.921
Macro	0.868	0.958	0.888	0.959	0.923	**0.875**

S5. Controlled architecture ablation

To characterize the contribution of each design choice, a controlled ablation was run around the selected architecture after it had been frozen, varying one factor at a time and training every variant from scratch under an identical short budget (five epochs, batch size 256, Adam at a learning rate of 10^{-3}, no early stopping) on QTDB (Table S5). Like the architecture search, the ablation placed windows at the annotated QRS onsets. Because the shared short budget under-trains all variants relative to the fully converged model, the absolute values are not comparable to the main results and should be read as *relative* effects. The picture is one of sharply diminishing returns in accuracy per parameter. Inserting an sLSTM block adds 1,390 parameters (a 55% increase) for +0.030 macro-F1 on LUDB, and enlarging the convolutional kernel to 21 or 41 adds 480 or 1,440 parameters for +0.019 and +0.046. A second attention head increases size while *reducing* LVAD-DB macro-F1 by 0.021, indicating no generalization benefit. Conversely, a shallower two-block stack (810 fewer parameters) reduces LUDB macro-F1 by 0.031. The selected model is thus the smallest configuration that remains on the accuracy–size frontier, consistent with its selection at the knee of the QTDB Pareto front. Because the ablation was evaluated after the selection, it describes the selected design and did not inform it. Bidirectionality is retained in all variants, as it is integral to the shared forward and backward stack.

Table S5: Controlled architecture ablation around the selected model (tolerance-aware macro-F1, LVAD-DB pooled across both raters). The first row gives the fully trained selected model (main-text Table I) for reference. In the remaining rows every variant was trained from scratch under an identical short budget (five epochs, batch size 256, Adam at a learning rate of 10^{-3}, no early stopping) on windows placed at the annotated QRS onsets and evaluated on windows placed in the same way. Because this budget under-trains all variants, the values should be read as *relative* effects. The Δ columns give the change in external macro-F1 relative to the short-budget selected model, computed from unrounded values.

Variant	Params	LUDB F1	LVAD-DB F1	ΔLUDB	ΔLVAD
Selected, fully trained (main-text Table I)	**2,522**	**0.919**	**0.875**	—	—
Controlled ablation, every variant trained from scratch under an identical short budget (Δ relative to the short-budget selected model)					
Selected (mLSTM, kernel 11, 1 head, 3 blocks)	2,522	0.837	0.847	—	—
+ sLSTM block	3,912	0.867	0.855	+0.030	+0.008
kernel size 21	3,002	0.856	0.861	+0.019	+0.014
kernel size 41	3,962	0.883	0.846	+0.046	−0.001
2 attention heads	2,816	0.848	0.827	+0.011	−0.021
2 blocks (shallower)	1,712	0.805	0.835	−0.031	−0.012

S6. Per-rater agreement on LVAD-DB

The main text reports LVAD-DB agreement pooled over the two independent experts, a biomedical engineer and a cardiologist. The cardiologist annotated excerpts from 26 recordings of 23 patients, and the engineer annotated five continuous 5-min segments from recordings of the same cohort. Signal matching linked three of the engineer's recordings to patient codes, one of them a patient also annotated by the cardiologist. The remaining two could not be linked to a patient code and were treated as separate resampling units in every patient-level bootstrap. Because these two recordings may belong to patients already represented, the pooled LVAD-DB confidence intervals rest on 27 resampling units for the 25 patients of the cohort. Table S6 gives the per-rater breakdown under strict and tolerance-aware evaluation, and Fig. S2 the per-rater confusion matrices, ROC curves, and boundary timing. Strict metrics use exact sample-wise matching, and tolerance-aware metrics apply the wave-specific matching windows ($\pm$52 ms P, $\pm$48 ms QRS, $\pm$124 ms T). AUC is

threshold-free and therefore identical under both regimes. Agreement is consistently higher against the biomedical engineer. Against the cardiologist, the gap is concentrated in six recordings in which wide, biphasic ventricular complexes were labelled as a single QRS complex (macro-F1 0.634 in these recordings vs. 0.858 in the other 20). With the alternative 8–18 Hz detector band the pooled tolerance-aware macro-F1 of the four models was Bi-mLSTM (this study) 0.878, Peimankar et al. 0.852, Jiménez-Pérez et al. 0.843, Liu et al. 0.869, in the same order as with the 5–15 Hz band.

Table S6: Per-rater agreement of the selected 2,522-parameter Bi-mLSTM model on LVAD-DB under strict and tolerance-aware evaluation. Rater 1 is the biomedical engineer (five recordings, 2,035 beat windows), Rater 2 the cardiologist (23 patients, 3,170 windows). All metrics are macro-averaged over the four classes. The second block repeats the evaluation with the alternative 8–18 Hz detector band. The lower block reports a stratified random subset of the cardiologist's beats (13 per patient) that the engineer re-annotated blind to the cardiologist's labels and to the model, after exclusion of 21 beats the engineer marked as uninterpretable. A second review by the engineer of the 65 beats in which the first-pass P-wave label overlapped the cardiologist's QRS label by at least 20 ms confirmed the first-pass labels. The first-named rater in each comparison is the test, the second the reference. Confidence intervals are from a patient-level bootstrap (2,000 resamples).

	Strict		**Tolerance-aware**		
Annotator	**F1**	κ	**F1**	κ	**AUC**
Detector band 5–15 Hz (primary analysis)					
Rater 1 (biomed. eng.)	0.823	0.751	0.938	0.918	0.962
Rater 2 (cardiologist)	0.710	0.605	0.824	0.765	0.897
Pooled	0.761	0.669	0.875	0.832	0.923
Alternative detector band 8–18 Hz (sensitivity analysis)					
Rater 1 (biomed. eng.)	0.823	0.751	0.938	0.918	0.962
Rater 2 (cardiologist)	0.719	0.616	0.829	0.772	0.898
Pooled	0.766	0.675	0.878	0.836	0.924

Beats annotated by both experts (278 beats, 23 patients; tolerance-aware)					
Comparison	**Macro-F1 [95% CI]**	κ **[95% CI]**	**P F1**	**QRS F1**	**T F1**
Engineer vs. cardiologist	0.784 [0.728, 0.829]	0.711 [0.648, 0.766]	0.725	0.846	0.833
Model vs. cardiologist	0.836 [0.795, 0.871]	0.781 [0.738, 0.822]	0.774	0.892	0.866
Model vs. engineer	0.884 [0.856, 0.909]	0.837 [0.799, 0.873]	0.882	0.923	0.894
Macro-F1 [95% CI], six wide-complex recordings (58 beats) / others (220 beats)					
Engineer vs. cardiologist	0.554 [0.475, 0.603] / 0.834 [0.804, 0.856]				
Model vs. cardiologist	0.662 [0.622, 0.679] / 0.866 [0.835, 0.891]				
Model vs. engineer	0.870 [0.831, 0.907] / 0.887 [0.854, 0.916]				

S7. Boundary timing and coverage

Each expert boundary was matched to the nearest predicted boundary of the same class and type within the class tolerance. Bias and SD are computed over matched boundaries only and are therefore conditional on a match. Table S7 reports, for every boundary type, the matched fraction of the expert boundaries, so that the excluded, unmatched boundaries are visible. Boundaries at the edges of contiguous window runs and at storage cuts are truncated rather than real and were removed before matching.

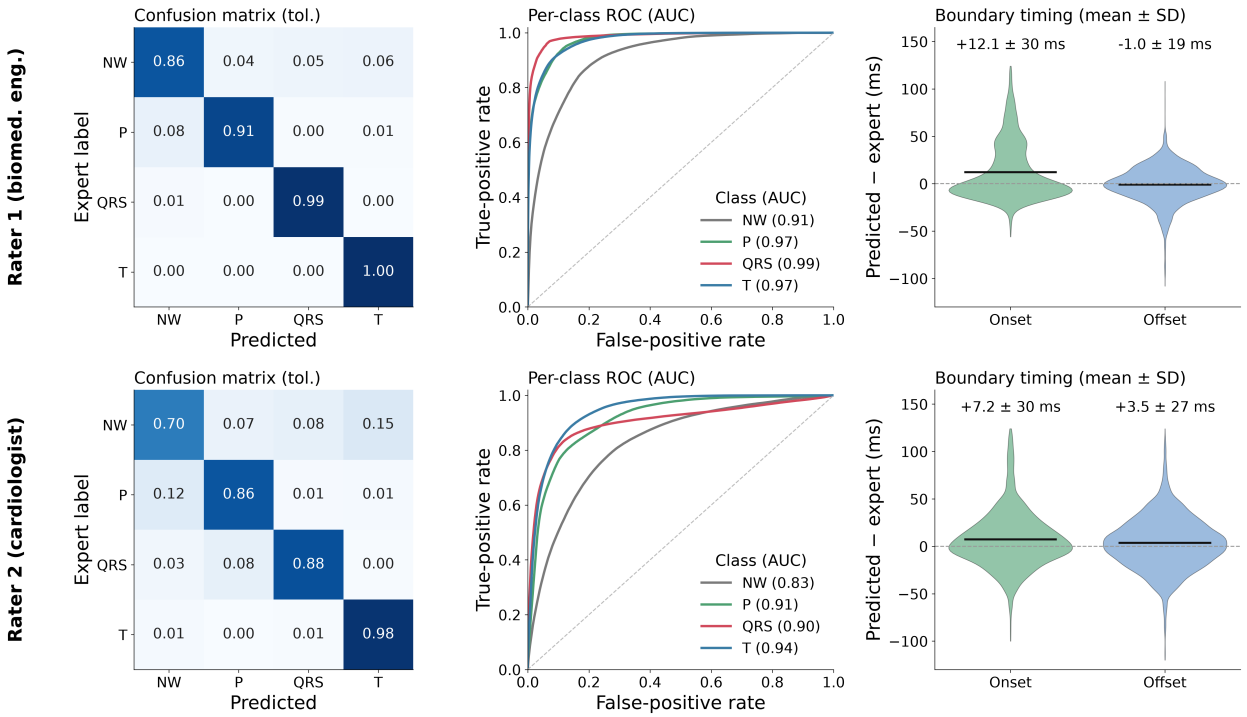


Figure S2: Per-rater delineation agreement of the selected 2,522-parameter Bi-mLSTM model on LVAD-DB for each expert (top row, Rater 1, the biomedical engineer, and bottom row, Rater 2, the cardiologist). **Left.** Row-normalized tolerance-aware confusion matrices over the four classes (NW, P, QRS, T). **Center.** Per-class one-vs-rest ROC curves with areas under the curve, with colors following the main text (P green, QRS red, T blue, NW gray). **Right.** Bland–Altman boundary timing, shown as distributions of model-minus-expert timing errors (ms) pooled over the onsets and over the offsets of P, QRS, and T, with the mean bias ± SD annotated above each violin and the zero-error line dashed. Boundaries were taken from the raw prediction on detector-placed windows and matched to the nearest expert boundary of the same class within the class tolerance, so the timing statistics are conditional on a match (Table S7). The bias ± SD was $+12.1 \pm 29.8$ ms (onsets) and -1.0 ± 18.9 ms (offsets) for Rater 1, and $+7.2 \pm 30.0$ ms and $+3.5 \pm 27.4$ ms for Rater 2. Agreement is higher against the biomedical engineer, with wider NW confusion and larger offset dispersion against the cardiologist.

Table S7: Boundary timing of the selected model on detector-placed windows (model − expert, ms) and matched fraction of the expert boundaries (Match., %). Bias ± SD over matched boundaries. LVAD-DB pooled across raters, with the matched fractions of the engineer (Eng.) and the cardiologist (Card.) given separately.

	LUDB		**LVAD-DB, pooled**		**Eng.**	**Card.**
Boundary	**Bias ± SD**	**Match.**	**Bias ± SD**	**Match.**	**Match.**	**Match.**
P onset	$+0.5 \pm 17.0$	90.3	$+5.2 \pm 18.7$	81.2	76.6	86.1
P offset	-0.8 ± 17.0	91.9	-1.8 ± 18.3	88.3	90.0	86.6
QRS onset	-6.6 ± 10.9	95.8	-3.8 ± 17.8	85.3	99.0	75.8
QRS offset	-1.2 ± 10.4	96.8	-0.9 ± 19.3	96.9	99.2	95.3
T onset	$+12.8 \pm 25.6$	98.5	$+25.9 \pm 37.7$	98.1	99.2	97.2
T offset	-1.1 ± 22.5	98.2	$+6.6 \pm 30.6$	99.0	99.8	98.3
All boundaries	$+0.6 \pm 19.1$	95.5	$+5.3 \pm 27.3$	91.7	94.0	89.8

S8. Clinical intervals

Table S8 extends main-text Table II with the availability of each interval in each source, bootstrap confidence intervals, per-record agreement, and per-rater values. Table S9 reports the agreement of model and expert on the normative-range classification in LUDB. LVAD-DB was not classified against normative ranges, because most patients with advanced heart failure on LVAD support have abnormal intervals. On LUDB, model-derived QRS duration and QT are systematically longer than the expert values, consistent with the early predicted QRS onset (Table S7), and this offset lowers the agreement on the QRS-duration normative classification.

Table S8: Agreement of model-derived and expert-derived clinical intervals in full (model − expert, ms). n, beats with the interval measurable in both sources (fraction of all beat windows). Avail. E / M, percentage of beats with the interval measurable from the expert and from the model labels. Bias ± SD and mean absolute error (MAE) per beat, with patient-level cluster-bootstrap 95% CIs (2,000 resamples; LVAD-DB pooled with both raters of a patient resampled together). Per record, bias ± SD of the difference between the model and expert medians of recordings with at least three measurable beats (number of recordings). For the biomedical engineer, the bootstrap clusters are the five continuously annotated recordings.

Interval	n (%)	Avail. E / M (%)	Bias ± SD [95% CI]	MAE [95% CI]	Per-record bias ± SD (n)
LUDB					
P duration	27,753 (78.0%)	84.4 / 87.5	−1.0 ± 34.0 [−2.9, +0.8]	24.4 [23.4, 25.5]	−1.9 ± 14.7 (171)
PR interval	27,753 (78.0%)	84.4 / 87.5	−14.6 ± 28.3 [−16.2, −13.1]	22.4 [21.3, 23.7]	−16.2 ± 12.5 (171)
QRS duration	34,521 (97.1%)	97.8 / 97.8	+12.8 ± 28.2 [+11.2, +14.2]	20.9 [19.9, 22.0]	+12.7 ± 11.8 (196)
QT interval	33,852 (95.2%)	97.1 / 96.1	+9.8 ± 43.8 [+7.7, +11.9]	28.4 [26.7, 30.2]	+11.3 ± 21.5 (196)
QTc (Bazett)	33,579 (94.4%)	95.9 / 95.1	+10.7 ± 47.6 [+8.3, +13.0]	30.8 [28.8, 33.1]	+12.8 ± 15.7 (194)
QTc (Fridericia)	33,579 (94.4%)	95.9 / 95.1	+10.4 ± 46.1 [+8.1, +12.7]	30.0 [28.0, 32.0]	+12.6 ± 15.0 (194)
LVAD-DB, pooled					
P duration	3,408 (65.5%)	68.9 / 83.9	−19.2 ± 38.1 [−32.7, −7.6]	29.9 [20.5, 41.3]	−19.8 ± 37.3 (25)
PR interval	3,408 (65.5%)	68.9 / 83.9	−20.3 ± 38.2 [−30.4, −12.7]	28.9 [21.8, 38.8]	−22.8 ± 31.2 (25)
QRS duration	4,241 (81.5%)	83.4 / 93.7	−3.4 ± 50.2 [−18.2, +8.4]	32.1 [25.1, 43.2]	−21.0 ± 67.5 (31)
QT interval	3,780 (72.6%)	74.5 / 86.7	+1.9 ± 61.8 [−15.5, +13.8]	39.0 [29.5, 54.5]	−23.0 ± 86.9 (28)
QTc (Bazett)	3,725 (71.6%)	72.4 / 84.0	+2.4 ± 73.3 [−16.5, +17.2]	46.9 [35.9, 64.2]	−27.3 ± 98.4 (28)
QTc (Fridericia)	3,725 (71.6%)	72.4 / 84.0	+2.1 ± 69.2 [−17.2, +15.6]	44.0 [33.4, 60.6]	−25.5 ± 95.0 (28)
LVAD-DB, biomedical engineer					
P duration	1,893 (93.0%)	98.9 / 93.6	−22.3 ± 35.8 [−43.4, −3.4]	29.6 [12.6, 46.9]	−19.2 ± 28.3 (5)
PR interval	1,893 (93.0%)	98.9 / 93.6	−15.4 ± 33.2 [−27.3, −7.3]	24.6 [13.8, 35.7]	−10.8 ± 14.4 (5)
QRS duration	2,026 (99.6%)	99.7 / 99.8	+5.3 ± 26.8 [−8.1, +21.2]	22.4 [17.9, 28.0]	+8.0 ± 20.2 (5)
QT interval	2,007 (98.6%)	98.7 / 99.6	+1.4 ± 35.8 [−9.3, +12.1]	24.2 [19.5, 27.7]	+4.8 ± 13.7 (5)
QTc (Bazett)	2,006 (98.6%)	98.6 / 99.5	+1.9 ± 43.0 [−10.5, +14.0]	28.7 [23.4, 33.1]	+6.3 ± 15.5 (5)
QTc (Fridericia)	2,006 (98.6%)	98.6 / 99.5	+1.7 ± 40.4 [−10.4, +13.2]	27.1 [22.1, 31.1]	+5.5 ± 14.3 (5)
LVAD-DB, cardiologist					
P duration	1,515 (47.8%)	49.6 / 77.8	−15.3 ± 40.5 [−27.7, −5.8]	30.2 [24.1, 40.3]	−19.9 ± 39.8 (20)
PR interval	1,515 (47.8%)	49.6 / 77.8	−26.3 ± 43.0 [−43.1, −14.5]	34.1 [24.6, 49.0]	−25.8 ± 33.7 (20)
QRS duration	2,215 (69.9%)	73.0 / 89.8	−11.4 ± 63.5 [−33.5, +4.6]	41.0 [29.5, 57.9]	−26.5 ± 72.1 (26)
QT interval	1,773 (55.9%)	58.9 / 78.4	+2.4 ± 81.9 [−30.0, +29.1]	55.7 [40.0, 78.5]	−29.0 ± 95.0 (23)
QTc (Bazett)	1,719 (54.2%)	55.5 / 74.1	+3.0 ± 97.5 [−35.8, +34.6]	68.1 [48.8, 95.2]	−34.6 ± 107.4 (23)
QTc (Fridericia)	1,719 (54.2%)	55.5 / 74.1	+2.5 ± 92.0 [−34.5, +33.9]	63.8 [45.7, 90.0]	−32.3 ± 103.7 (23)

S9. Calibration

The ECE is a top-label ECE over all non-padded samples of a dataset, pooled across classes, patients, and windows. Confidence is the maximum softmax probability, and the interval $[0, 1]$ is divided into 15 equal-width bins with the top bin closed. For each non-empty bin b the absolute difference between the fraction of correct samples and the mean confidence is weighted by the fraction of samples in the bin, $\mathrm{ECE} = \sum_b (n_b/N)\,|\mathrm{acc}_b - \mathrm{conf}_b|$, and empty bins contribute nothing. Confidence intervals resample patients and recompute the ECE from the summed per-patient bin

Table S9: Agreement of model and expert on whether intervals lie within the normative range, in LUDB records in sinus rhythm without pacing (176 records, of which 170 to 172 had at least three measurable beats). QTc thresholds are 450 ms for men and 460 ms for women, with sex taken from the LUDB headers. Per beat, the fraction of beats within range for expert and model and Cohen's κ. Per record, the record medians were classified and compared (agreement and κ).

	Per beat				Per record		
Interval (normative range)	**Beats**	**Expert in range**	**Model in range**	κ	**Records**	**Agreement**	κ
P duration (≤120 ms)	27,264	87.4%	80.8%	0.29	170	97.1%	0.65
PR interval (120–200 ms)	27,264	84.7%	75.9%	0.24	170	80.6%	0.09
QRS duration (≤120 ms)	29,823	84.3%	70.8%	0.47	172	88.9%	0.35
QTc, Bazett (<450/460 ms)	29,165	79.7%	71.2%	0.56	170	90.0%	0.58
QTc, Fridericia (<450/460 ms)	29,165	84.5%	77.9%	0.55	170	92.9%	0.47

counts. The reliability diagrams use the same bins and show only bins with at least 50 samples. Against the original labels a sample counts as correct when the argmax prediction equals the reference label. Against the tolerance-matched labels it counts as correct when the tolerance-matched prediction equals the reference label, at unchanged confidence. Tolerance matching relabels only predictions at annotated wave samples, and only towards the annotation, when the same class was predicted within the class tolerance in the same window. It never changes samples annotated as NW or spurious wave predictions. On LUDB it relabelled 5.7% of the scored samples, mostly low-confidence boundary samples, which then count as correct at their original confidence. The model therefore appears under-confident against tolerance-matched labels. For temperature scaling, one scalar T was fitted by minimizing the negative log-likelihood (NLL) over the non-padded samples of the 21 QTDB validation records ($T = 0.714$, validation NLL 0.354 to 0.325) and applied unchanged to both external datasets (Table S10). Because $T < 1$ sharpens the softmax, it corrected the under-confidence on LUDB but made the model over-confident on LVAD-DB. For reference, the table also gives the temperature that each external dataset would select for itself. This value uses the test labels and is not an independent estimate.

Table S10: Expected calibration error (ECE) of the selected model before and after temperature scaling with the temperature fitted on the QTDB validation records ($T = 0.714$), against the original and the tolerance-matched labels, with patient-level bootstrap 95% CIs. NLL, negative log-likelihood before / after scaling. Units, patient-level resampling units. Same-set T, the temperature fitted on each external set itself, given for reference only.

		ECE, original labels		ECE, tolerance-matched labels			
Dataset	**Units**	**Raw**	**After QTDB T**	**Raw**	**After QTDB T**	**NLL**	**Same-set T**
LUDB	199	0.054 [0.050, 0.058]	0.013 [0.010, 0.017]	0.111 [0.107, 0.114]	0.044 [0.040, 0.048]	0.349 / 0.321	0.72
LVAD-DB, pooled	27	0.041 [0.018, 0.070]	0.107 [0.084, 0.137]	0.091 [0.071, 0.111]	0.068 [0.056, 0.086]	0.645 / 0.722	1.13
LVAD-DB, engineer	5	0.019 [0.016, 0.042]	0.059 [0.030, 0.088]	0.129 [0.105, 0.154]	0.064 [0.040, 0.089]	0.470 / 0.485	0.90
LVAD-DB, cardiologist	23	0.076 [0.049, 0.103]	0.144 [0.117, 0.171]	0.078 [0.062, 0.103]	0.088 [0.069, 0.110]	0.777 / 0.900	1.29

S10. Comparison with published architectures

Table S11 gives the macro-averaged metrics of the proposed model and the three baselines, all trained with the same windowing and procedure, under strict and tolerance-aware evaluation. Table S12 gives the paired differences with patient-level bootstrap confidence intervals and Holm-adjusted p-values. Under tolerance-aware scoring our model exceeds every baseline on LUDB, and its margin over Liu et al. on LVAD-DB is not significant. Under strict scoring Peimankar et al. and

Liu et al. exceed our model on LUDB, mainly through a higher P-wave F1, and no macro-F1 difference on LVAD-DB is significant after correction. Table S13 reports the pipeline-level comparison, in which each baseline was run in its native, published pipeline with its original checkpoint, trained and evaluated on unanchored, non-overlapping 1.2-s sliding windows (300 samples, zero-padded tail, per-window z-normalization, same 0.5–40 Hz band-pass). The baselines trained and evaluated on beat-anchored windows (main-text Table III) scored higher than in their native pipelines, by +0.007 to +0.017 macro-F1 on LUDB and +0.116 to +0.145 on LVAD-DB. Because the native pipelines also differ in checkpoint, training data, and preprocessing, these differences are descriptive and do not isolate the effect of anchoring.

Table S11: Macro-averaged diagnostic metrics for the proposed model and the three baselines on both external datasets under strict and tolerance-aware evaluation (main-text Table III). All four models were trained from scratch with the same windowing, augmentation, loss, optimizer, schedule, split, and stopping rule, and evaluated with the same pipeline. Metrics are macro-averaged over the four classes (NW, P, QRS, T). κ is Cohen's kappa, and AUC is the mean one-vs-rest area under the ROC curve. LVAD-DB is pooled across both expert raters. Paired differences are given in Table S12.

Model	PPV	NPV	Sens.	Spec.	AUC	κ	F1
LUDB, strict							
Bi-mLSTM (this study)	0.848	0.946	0.855	0.946	0.977	0.793	**0.850**
Peimankar et al.	0.851	0.946	0.884	0.951	0.970	0.804	0.867
Jiménez-Pérez et al.	0.845	0.942	0.853	0.945	0.964	0.785	0.848
Liu et al.	0.857	0.949	0.883	0.953	0.977	0.813	0.869
LUDB, tolerance-aware							
Bi-mLSTM (this study)	0.891	0.966	0.954	0.978	0.977	0.891	**0.919**
Peimankar et al.	0.880	0.960	0.951	0.974	0.970	0.871	0.911
Jiménez-Pérez et al.	0.875	0.957	0.922	0.968	0.964	0.855	0.896
Liu et al.	0.887	0.964	0.951	0.976	0.977	0.882	0.915
LVAD-DB (pooled), strict							
Bi-mLSTM (this study)	0.769	0.916	0.754	0.915	0.923	0.669	**0.761**
Peimankar et al.	0.746	0.906	0.734	0.905	0.896	0.630	0.738
Jiménez-Pérez et al.	0.740	0.902	0.713	0.900	0.897	0.613	0.724
Liu et al.	0.762	0.915	0.757	0.915	0.914	0.667	0.758
LVAD-DB (pooled), tolerance-aware							
Bi-mLSTM (this study)	0.868	0.958	0.888	0.959	0.923	0.832	**0.875**
Peimankar et al.	0.839	0.947	0.867	0.948	0.896	0.792	0.849
Jiménez-Pérez et al.	0.834	0.944	0.849	0.945	0.897	0.779	0.839
Liu et al.	0.859	0.955	0.885	0.957	0.914	0.821	0.866

S11. On-device feasibility of baseline architectures

Table S14 compares the on-device memory footprint of the proposed model with that of the baselines. The compact model is the only architecture whose weights fit within the microcontroller's 1 MB on-chip SRAM at the deployed 32-bit precision. The baselines were not deployed, so their rows indicate that they would require external memory or quantization on this device, not that they cannot run on a microcontroller.

Table S12: Paired differences between the proposed model and each baseline (ours − baseline) in macro-F1 and Cohen's κ, with 95% CIs from a paired patient-level bootstrap (2,000 resamples; both models scored on the same resampled patients) and two-sided bootstrap p-values adjusted with the Holm method over three baselines and two datasets for each metric and scoring regime (p_{Holm}). The smallest attainable adjusted value with 2,000 resamples is 0.006.

Baseline	**Δ macro-F1 [95% CI]**	p_{Holm}	**Δκ [95% CI]**	p_{Holm}
LUDB, tolerance-aware				
Peimankar et al.	+0.008 [+0.005, +0.011]	<0.01	+0.020 [+0.015, +0.025]	<0.01
Jiménez-Pérez et al.	+0.023 [+0.020, +0.026]	<0.01	+0.036 [+0.032, +0.040]	<0.01
Liu et al.	+0.004 [+0.002, +0.006]	<0.01	+0.009 [+0.006, +0.012]	<0.01
LUDB, strict				
Peimankar et al.	−0.016 [−0.020, −0.013]	<0.01	−0.011 [−0.016, −0.006]	<0.01
Jiménez-Pérez et al.	+0.002 [−0.001, +0.006]	0.376	+0.008 [+0.004, +0.013]	<0.01
Liu et al.	−0.019 [−0.022, −0.017]	<0.01	−0.020 [−0.023, −0.016]	<0.01
LVAD-DB, tolerance-aware				
Peimankar et al.	+0.025 [+0.006, +0.045]	0.026	+0.040 [+0.013, +0.069]	0.010
Jiménez-Pérez et al.	+0.035 [+0.016, +0.053]	<0.01	+0.052 [+0.025, +0.082]	<0.01
Liu et al.	+0.008 [−0.001, +0.018]	0.067	+0.011 [−0.001, +0.024]	0.079
LVAD-DB, strict				
Peimankar et al.	+0.023 [+0.000, +0.051]	0.141	+0.040 [+0.005, +0.080]	0.038
Jiménez-Pérez et al.	+0.037 [+0.008, +0.067]	0.052	+0.057 [+0.015, +0.105]	0.018
Liu et al.	+0.002 [−0.012, +0.020]	0.807	+0.002 [−0.019, +0.028]	0.847

Table S13: Pipeline-level comparison (tolerance-aware macro-F1, LVAD-DB pooled across both raters). Each baseline in its native pipeline (original checkpoint on unanchored, non-overlapping 1.2-s sliding windows) and our model in its beat-anchored pipeline and on the same unanchored windows. The native pipelines differ from ours in checkpoint, training data, and preprocessing as well as in windowing, so this table is descriptive. Weight footprints are for 32-bit floating point (parameters × 4 bytes).

Model (Ref.)	**Params (k)**	**Input windowing**	**LUDB F1**	**LVAD-DB F1**	**Weights (float32)**
Bi-mLSTM (this study)	**2.5**	beat-anchored (detector-placed)	**0.919**	**0.875**	**10 kB**
Bi-mLSTM (this study)	2.5	unanchored sliding windows	0.834	0.653	10 kB
Peimankar et al.	1,419	unanchored sliding windows (native)	0.903	0.723	5.7 MB
Jiménez-Pérez et al.	1,296	unanchored sliding windows (native)	0.889	0.723	5.2 MB
Liu et al.	301	unanchored sliding windows (native)	0.899	0.721	1.2 MB

Table S14: Deployment feasibility on microcontroller-class edge hardware (Arduino Portenta H7). Weight footprints are given for all models in 32-bit floating point (4 B per parameter), the precision in which the proposed model is deployed and the one shown in main-text Fig. 4d, with the best-case INT8 size (1 B per parameter) in parentheses (1 kB = 10^3 B, 1 MB = 10^6 B). Peak activation memory is estimated from each architecture's largest intermediate tensor. "Fits on-chip SRAM?" compares the total (weights, activations, and runtime and IO buffers) against the device's usable on-chip SRAM and flash budget. Only the proposed model fits in 32-bit floating point without quantization, since the float32 weights of all three baselines alone (1.2–5.7 MB) exceed the 1 MB on-chip SRAM. After INT8 quantization the Peimankar et al. and Jiménez-Pérez et al. weights (1.42 and 1.30 MB) still exceed it, whereas the Liu et al. weights (0.30 MB) would fit, at the cost of the quantization that the proposed model does not need. On-device values for the proposed model (peak heap SRAM and per-beat latency) were measured on the Arduino Portenta H7 (Cortex-M7, 1 MB on-chip SRAM, 2 MB flash) over the 469-beat pilot recording. The baseline rows are architecture-based estimates, as the baselines were not deployed.

Model	Params	Weights, float32 (INT8)	Est. peak activations	On-device latency	Fits on-chip SRAM?
Bi-mLSTM (this study)	2.5 k	10 kB (2.5 kB)	22 kB (heap)	128.6 ± 1.5 ms	**Yes** (deployed, float32)
Peimankar et al.	1.4 M	5.7 MB (1.42 MB)	$\gg$ budget	—	No, weights exceed SRAM in float32 and INT8
Jiménez-Pérez et al.	1.3 M	5.2 MB (1.30 MB)	$\gg$ budget	—	No, weights exceed SRAM in float32 and INT8
Liu et al.	301 k	1.2 MB (0.30 MB)	$\gg$ budget	—	No in float32; INT8 weights would fit

S12. Firmware verification and deployment timing

To verify the firmware, the pilot recording was replayed offline through a pure-NumPy recurrent-form reimplementation of the firmware classifier. The replay received the device-logged beat windows and the filtered, chunk-normalized input (zero-padded to 300 samples), and outputs were stitched exactly as in the firmware. With the weights that ran on the device, the replay reproduced the device's logged per-sample labels exactly (0 of 90,306 samples differing, 469 of 469 windows identical), and the NumPy and PyTorch implementations agreed to within 4.3×10^{-6} in class probability. The pilot firmware ran an earlier set of weights of the identical architecture. Feeding the same windows to the final weights of the main text therefore emulates their on-device output (Table S15). Differences were scored with the same tolerance-aware pipeline as the device output, and 95% confidence intervals are from a paired bootstrap over the 73 annotated beats (2,000 resamples, both raters resampled jointly). The values for the final weights are an emulation of the unchanged firmware, not an on-device measurement. The same replay with the offline normalization convention, in which each zero-padded window is z-normalized individually instead of taking the windows from the buffer-normalized signal, changed the pooled macro-F1 by -0.004 [-0.007, -0.000] for the pilot weights and by -0.009 [-0.018, -0.003] for the final weights. The normalization difference between device and offline evaluation therefore changes the pilot macro-F1 by less than 0.01.

The firmware classifies a beat as soon as the following R peak has been detected, because the bidirectional model needs the complete window up to the next anchor. The labels of a beat are therefore available about one RR interval after its onset, plus the R-peak confirmation and the processing time of 128.6 ms (preprocessing 1.5 ms, inference 127.1 ms). At the median RR interval of the pilot recording (0.77 s) this amounts to about 0.9 s. BLE transport and output streaming are not included. Two of the 469 pilot windows (0.4%) exceeded 300 samples and were truncated on the device, whereas such intervals were excluded in the offline evaluation.

Table S15: Firmware verification and emulation on the pilot recording (single healthy subject, 361 s, five annotated 10-s spans, two raters, tolerance-aware). *Pilot weights*, replay of the weights that ran on the device, identical to the logged device output. *Final weights*, the same firmware path with the final weights of the main text, an emulation rather than an on-device measurement. Δ = final − pilot, with paired-bootstrap 95% CI over 73 annotated beats.

Metric (tolerance-aware)	**Pilot weights**	**Final weights**	**Δ [95% CI]**
Pooled macro-F1	0.926	0.949	+0.022 [+0.014, +0.033]
Pooled κ	0.893	0.925	+0.032 [+0.021, +0.047]
Macro-F1, engineer	0.929	0.955	+0.026 [+0.017, +0.038]
Macro-F1, cardiologist	0.923	0.942	+0.018 [+0.010, +0.030]
F1, NW	0.916	0.940	+0.025 [+0.017, +0.032]
F1, P wave	0.951	0.937	−0.014 [−0.022, −0.006]
F1, QRS	0.890	0.951	+0.061 [+0.042, +0.085]
F1, T wave	0.949	0.965	+0.016 [+0.003, +0.035]
Macro-F1, rest 1	0.954	0.981	+0.027 [+0.021, +0.034]
Macro-F1, walking	0.926	0.943	+0.018 [+0.006, +0.027]
Macro-F1, rest 2	0.912	0.937	+0.024 [+0.008, +0.052]
Offline normalization convention (per-window z-normalization)			
Pooled macro-F1	0.922	0.939	—
Δ vs. device normalization[a]	−0.004	−0.009	—

[a]Paired-bootstrap 95% CIs $[-0.007, -0.000]$ (pilot weights) and $[-0.018, -0.003]$ (final weights), 73 annotated beats.