

Article

QTANet: Boundary-Aware Deep Learning for Single-Lead QT-Interval Measurement

Sang-Bum Park ¹, Myungeun Lee ², Jinwon Heo ¹, Yoo Ri Kim ³ and Hyung-Jeong Yang ^{1,2,4,*}

¹ Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, Republic of Korea; sb8919@jnu.ac.kr (S.-B.P.); jwww0124@jnu.ac.kr (J.H.)

² Hyper-Wide Federated Medical AI Research Center, Chonnam National University, Gwangju 61186, Republic of Korea; melee@jnu.ac.kr

³ Department of Cardiology, Chonnam National University Hospital, Gwangju 61469, Republic of Korea; yoorikim@jnu.ac.kr

⁴ The Artificial Intelligence Convergence Research Institute, Chonnam National University, Gwangju 61186, Republic of Korea

* Correspondence: hjyang@jnu.ac.kr

Featured Application

Single-lead ECG measurement of the QT interval with inspectable QRS-onset and T-wave-offset landmarks, intended to support research on automated cardiac monitoring and clinician review.

Abstract

Accurate QT-interval measurement requires reliable identification of QRS onset and T-wave offset. We present QTANet, a single-lead framework that classifies ECG samples as baseline, QRS or T wave and derives QT from the detected landmarks. We used 179 recordings from the Lobachevsky University Electrocardiography Database, with a fixed test set and patient-wise cross-validation, and external evaluation without fine-tuning on the QT Database (QTDB) and 33 clinical recordings. On the expert-annotated QTDB interval subset of 101 recordings, QTANet achieved a mean absolute error of 34.94 ms and measurement coverage of 95.2%, compared with 40.55 ms and 88.2% for the control without attention and 52.41 ms and 68.7% for NeuroKit2. In a separate expert-annotated subset permitting frame-level evaluation, attention increased per-recording T-wave F1 by 0.061, with improvement in 30 of 41 recordings. Attention most consistently reduced T waves misread as QRS complexes. QTANet outperformed an identically trained published U-Net externally at frame level but had higher internal QT error; a larger modern U-Net outperformed QTANet at frame level and in QTDB QT error. Internal interval-level gains were small and clinical results varied across training seeds. These findings support boundary-derived QT measurement with jointly reported error and coverage; clinical utility requires further validation.



Academic Editor: Emi Yuda

Received: 7 September 2026

Revised: 7 October 2026

Accepted: 13 October 2026

Published: 17 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: electrocardiogram; QT interval; ECG delineation; QRS onset; T-wave offset; discrete wavelet transform; self-attention; distribution shift

1. Introduction

Accurate interpretation of the electrocardiogram (ECG) is fundamental to cardiovascular diagnosis [1]. The QT interval—the time from QRS-complex onset to T-wave

offset—quantifies ventricular depolarization and repolarization and is central to arrhythmia-risk assessment [2]. Prolongation of the heart-rate-corrected QT interval (QTc) is a recognized marker of torsades-de-pointes risk and drug-induced cardiotoxicity, which makes QT/QTc measurement safety-critical in the clinical monitoring of patients receiving QT-prolonging drugs [3] and in regulatory drug-safety assessment [4]. Because risk stratification acts on QTc, any error in locating QRS onset or T-wave offset propagates directly to the reported QTc and to patient safety.

Reliable QT measurement is difficult. Manual measurement shows substantial disagreement even among experts, particularly when identifying the end of notched, bifid or low-amplitude T waves and distinguishing T from U waves [2]. In patients with long-QT syndrome, QTc values measured by expert readers have been reported to vary by up to 70 ms [5]. Differences in lead configuration, sampling and patient population further complicate transfer between databases. Classical rule-based delineators [6], including wavelet-based implementations available in NeuroKit2 [7], provide a relevant comparison for both measurement error and the proportion of beats for which an interval is returned (Section 3.5).

Deep learning has advanced automated ECG analysis through data-driven feature learning [1,8–10]. QTNet estimates QT from lead I using a multi-task model trained on more than three million device-labeled ECGs [11,12], whereas QTcNet estimates heart-rate-corrected QT using an InceptionTime backbone [13]. Scalar prediction provides an interval without explicitly locating the boundaries that define it. Post-hoc attribution can indicate influential waveform regions, but does not directly provide the landmarks used to calculate the interval. We therefore examine boundary-derived QT estimation and compare it with scalar regression under the same limited training-data setting (Section 3.2).

Boundary-aware delineation classifies samples as baseline, QRS or T wave and derives QT from detected landmarks. This provides output-level traceability: a clinician can inspect the boundaries associated with a reported interval. The present study evaluates measurement accuracy and coverage, the contribution of self-attention and its effect on waveform-identity confusion and false-positive segments, a comparison with a published deep-learning delineator under an identical protocol, and performance on external recordings with heterogeneous acquisition characteristics. Boundary visibility is treated as a property of the output rather than evidence, by itself, of clinical usefulness.

QTANet couples a residual convolutional encoder, a discrete-wavelet branch with squeeze-and-excitation (SE) channel weighting, and two bidirectional LSTM stages separated by a self-attention block. A class-weighted focal loss uses a scheduled focusing parameter. The study uses LUDB [14] for model development and internal evaluation, QTDB [15] for external evaluation against manual and automatic references, and an independent clinical recording cohort from CNUH. The annotation standard, evaluation subset and aggregation unit are distinguished when interpreting the results.

1.1. Related Work

Classical delineation pipelines use digital filters, wavelet features and threshold rules to locate fiducial points [6]; NeuroKit2 provides a reproducible implementation framework [7]. Some beats do not yield a valid interval, so an MAE calculated only on successful measurements is conditional on a method's coverage. We report coverage alongside the main beat-level QT results and provide a descriptive analysis of the subset measured by NeuroKit2 (Section 3.5).

Deep-learning delineators built on one-dimensional U-Net encoder–decoders [16,17] and their residual and recurrent variants [18] achieve high per-wave agreement on curated databases under a tolerance-based (e.g., 150 ms) detection protocol, and models trained on

PTB-XL have also been validated against the independently annotated CSE database [18]. Transformer encoders have also been applied to ECG classification [10] and benchmarked for semi-supervised delineation on LUDB and QTDB [19], and state-space (Mamba) layers have been used for synchronous P, QRS and T delineation in 12-lead ECGs [20]. Reported “F1 > 99%” figures use this wave-detection definition, which differs from frame-level classification F1; we report both to remain comparable (Section 3.1). Recent work also shows that models trained only on public databases can degrade on under-represented rhythms [16], underscoring the importance of external, distribution-shifted evaluation. These studies report false positives as aggregate counts within sensitivity and positive predictive value. To our knowledge, where spurious segments fall relative to the reference waves, which distinguishes confusions of wave identity from segments overlapping no annotated wave, has not been analyzed; we introduce such a decomposition (Section 3.10).

Regression-based systems estimate a scalar QT or QTc value from an ECG. QTNet uses lead I and large-scale device-derived labels [11,12]. QTcNet uses 12-lead inputs and targets QTc; its Integrated Gradients analysis identifies influential waveform regions [13]. Our regression comparison uses the same recording pool and fold split as QTANet at a much smaller annotation scale. It evaluates the implemented formulations under that training budget and does not establish the relative performance of systems trained on millions of recordings.

Explainability in ECG AI is dominated by post-hoc attribution (Grad-CAM, Integrated Gradients, SHAP) [21]. Boundary-aware delineation offers complementary, output-level transparency: because the reported QT is a function of detected landmarks, the measurement is directly verifiable, without a separate attribution step.

1.2. Contributions

QTANet is assembled from established components (residual convolution, discrete wavelet transform, bidirectional LSTM and multi-head self-attention), and boundary-aware delineation is itself an established approach; we do not claim novelty for the individual modules. The task-specific design is listed first and is distinguished from the analytical and empirical contributions that follow.

1. **Task-specific design.** For single-lead QT delineation, QTANet combines morphological and wavelet features, processes the fused representation with a first Bi-LSTM, applies self-attention to contextualize the temporal features and passes the result through a second Bi-LSTM before sample-wise classification (Sections 2.3–2.6). The rationale, that waveform identification may benefit from information beyond the local shape of an individual deflection, is evaluated in multi-seed component ablations and a sensitivity analysis of the attention placement (Section 3.7) rather than assumed, and the configuration is not claimed to be optimal.
2. **Analysis of waveform-identity errors.** We identify a failure mode of single-lead delineation under distribution shift. On QTDB, the model without attention generates many spurious waves (QRS PPV 82.73%), and a large share of them are confusions of wave identity, with T waves labeled as QRS complexes and P waves labeled as T waves (Sections 3.1 and 3.10).
3. **Error decomposition.** We introduce a decomposition of false-positive segments by the reference wave they overlap, which makes this failure mode measurable. Across four training seeds, adding self-attention is associated with 55.0–90.7% fewer QRS segments predicted over reference T waves, a larger reduction in every seed than for segments overlapping no annotated wave (21.3–55.1%), whereas the change for T-wave segments over reference P waves is not consistent across seeds. Recordings with larger reductions in wave-identity confusion show larger gains in measurement

coverage. Against manual cardiologist annotations, self-attention improves T-wave F1 in 30 of 41 fully bracketed QTDB recordings (Sections 3.6 and 3.10).

4. **Matched-protocol comparison.** We compare QTANet with the published one-dimensional U-Net delineator of Moskalenko et al. [17], reimplemented to its specification and trained under an identical protocol. QTANet is superior in frame-level metrics on both external cohorts and in QT MAE on QTDB, whereas on the internal out-of-fold data the U-Net has higher T-wave F1 and lower QT MAE; a larger modern U-Net variant outperforms QTANet on frame-level metrics and, averaged over four seeds, in QT MAE on QTDB. The benefit is therefore observed under distribution shift, relative to a much smaller published delineator, and does not indicate architectural optimality (Section 3.11).
5. **Interval accuracy and coverage.** We report boundary-derived QT measurement together with measurement coverage under internal and external protocols. On the expert-annotated QTDB interval subset, QTANet has lower QT MAE than the control without attention, ResNet + LSTM + DWT and NeuroKit2, and across the reported dataset conditions it returns valid intervals for 92.8–97.1% of reference beats, compared with 68.2–90.8% for NeuroKit2 (Sections 3.3 and 3.5). A separate window-level comparison with scalar regression is confined to a limited expert-annotation budget (Section 3.2).
6. **Transparency and limits.** We expose the waveform landmarks used to derive each QT interval and report the limits of the evaluation. The analyses distinguish expert from automatic references, fixed-test from out-of-fold evaluation, and variation across recordings from variation across training seeds. Internal interval-level gains and the consistency of clinical effects are limited.

The remainder of the paper is organized as follows. Section 2 describes QTANet together with the datasets, baselines, and evaluation protocol; Section 3 reports the results; Section 4 discusses the findings and limitations; Section 5 concludes.

2. Materials and Methods

2.1. Problem Formulation and Overview

The QT interval is the temporal distance between QRS-complex onset and T-wave offset. Rather than regressing a single value, QTANet formulates QT estimation as boundary-aware delineation: for a fixed-length ECG segment it produces, for every sample i , class probabilities $p_{i,c}$ over $c \in \{\text{baseline}, \text{QRS}, \text{T}\}$, and the QT interval is derived from the detected onset/offset. An overview is shown in Figure 1. A residual convolutional encoder and a parallel discrete-wavelet branch produce complementary morphological and low-frequency features; these are fused and passed through a bidirectional LSTM stage, a self-attention block with layer normalization before the attention sublayer and after the feed-forward sublayer, and a second bidirectional LSTM stage; a classifier emits per-sample class probabilities; and a beat-wise step derives QRS onset and T-wave offset and hence QT.

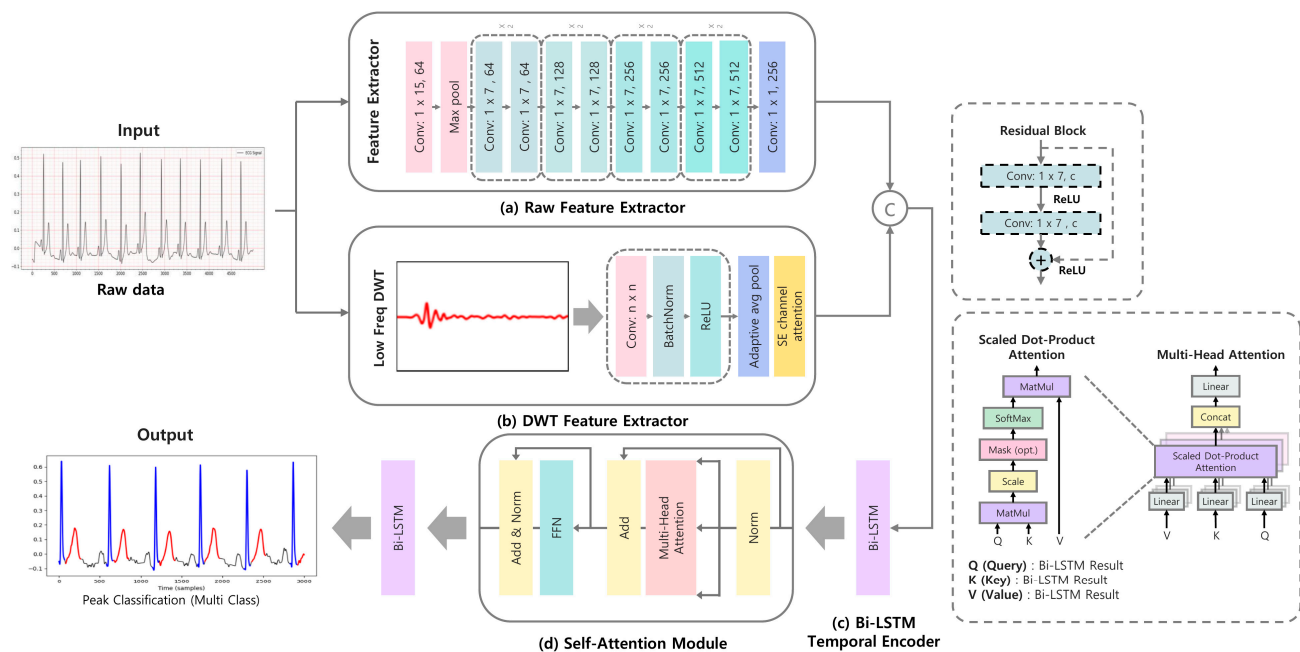


Figure 1. Architecture of QTANet. Residual convolutional and discrete-wavelet features are combined with bidirectional LSTM stages and self-attention for sample-wise classification into baseline, QRS and T wave. Sections 2.3–2.6 describe the component operations and QT derivation. Arrows indicate the data flow, and $\times 2$ marks residual stages of two blocks; in the output, blue and red mark samples classified as QRS complex and T wave, respectively.

2.2. Preprocessing and Frame-Level Labeling

ECGs are preprocessed with the default method of the `ecg_clean` function of `NeuroKit2` (v0.2.12) [7], which applies two zero-phase stages. Baseline wander is removed with a fifth-order high-pass Butterworth filter with a cut-off frequency of 0.5 Hz, implemented in second-order sections and applied forward and backward (`scipy.signal.sosfiltfilt`; SciPy v1.17.1). Power-line interference is then suppressed with a moving-average kernel one 50 Hz period wide (10 samples at 500 Hz), also applied forward and backward (`scipy.signal.filtfilt`). Both filtering stages were applied in a zero-phase manner to avoid phase delay; however, filtering may still alter waveform morphology and consequently affect estimated fiducial locations. No further filtering is applied. QTDB, distributed at 250 Hz, is resampled to 500 Hz by Fourier-domain resampling (`scipy.signal.resample`); because the ratio is exactly two, annotation sample indices are mapped by multiplication by two, without interpolation or rounding. Resampling and index mapping precede filtering. LUDB and CNUH are natively sampled at 500 Hz and are not resampled; CNUH signals, stored in device ADC units, are converted to millivolts (4.88 μV per least significant bit) at load time. Fixed 6-s windows are then extracted (Section 2.8). Samples are labeled baseline, QRS or T using the available waveform boundaries (Figure 2). LUDB and CNUH use expert annotations. QTDB uses manual annotations for the specified expert-reference subsets and automatic annotations for full-coverage analyses (Section 2.8). Within each analyzed window, samples outside the annotated QRS and T-wave intervals, including P-wave samples, are assigned to the baseline class, and no samples are discarded as unlabeled; the unannotated margins of the 10-s LUDB and CNUH recordings lie outside the analyzed window (Section 2.8.1). Because frame labels require both the onset and the offset of every wave, expert-reference frame metrics on QTDB are restricted to the recordings in which every T wave is fully bracketed (Section 2.8.2). The baseline class thus denotes samples outside the annotated QRS and T-wave intervals rather than expert-confirmed baseline.

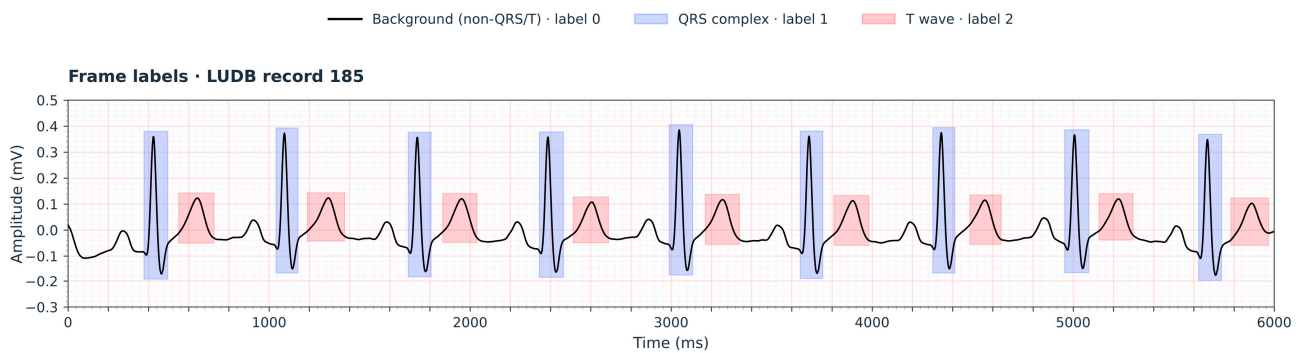


Figure 2. Three-class frame labeling. Each sample is assigned to baseline (non-QRS/T), QRS complex or T wave using the available waveform boundaries. The shaded regions illustrate the two waveform classes.

The 6-s window length follows from the structure of the two 10-s datasets and was not optimized. In LUDB and CNUH, the central 2–8 s of each recording is analyzed; discarding the first and last 2 s removes electrode-contact transients and lead-switching artifacts, which occur predominantly at the start and end of a short resting acquisition, as well as the edge distortion that zero-phase filtering introduces at segment boundaries. The 15-min QTDB recordings are tiled into consecutive windows of the same length, so that the input length is identical across the three cohorts and the external comparison is not confounded by window length. At 60–100 beats per minute, a 6-s window contains 6–10 beats, which provide rhythm context for the recurrent and attention stages. Window length was not varied empirically, because the unannotated margins of the 10-s recordings leave a narrow usable range; it is a design decision and was not tuned (Section 4.4).

2.3. Residual Convolutional Encoder

Morphological features are extracted by a one-dimensional residual convolutional encoder [22]. A residual block computes

$$h^{(l+1)} = \text{ReLU} \left(\mathcal{F} \left(h^{(l)}; \theta_l \right) + h^{(l)} \right) \quad (1)$$

with $\mathcal{F}$ a stack of 1-D convolutions with batch normalization: two 1×7 convolutions, each followed by batch normalization, with ReLU and dropout (0.2) between them, and a 1×1 convolution with batch normalization on the shortcut where the number of channels changes. The raw-signal branch consists of an initial convolution (kernel size 15, 64 channels), a max-pooling layer with kernel size 3 and stride 1, four residual stages of two blocks each with 64, 128, 256 and 512 channels (the ResNet-18 layout), and a 1×1 convolution that projects the 512 channels to 256 (Figure 1). All convolutions and the pooling layer use stride 1 with length-preserving padding, so no layer reduces the temporal resolution, and the encoder output has 256 channels at the full input length of 3000 samples. The parallel DWT branch contributes 128 channels at the same length (Section 2.4), so that the fused representation has 384 features per time step.

2.4. Discrete-Wavelet Branch with Channel Attention

A parallel wavelet branch supplies frequency-resolved features that complement the morphological features of the convolutional encoder. Its contribution is evaluated through component ablation rather than assumed to confer noise robustness in every setting.

Using a Daubechies-8 (db8) basis, the DWT decomposes the segment via the Mallat recursion,

$$a_j[n] = \sum_k h[k - 2n] a_{j-1}[k], \quad d_j[n] = \sum_k g[k - 2n] a_{j-1}[k], \quad a_0 = x \quad (2)$$

Here h and g denote the low- and high-pass filters, respectively. The approximation and detail coefficients follow Equation (2). The decomposition uses PyWavelets (v1.8.0) with its default signal-extension mode, “symmetric”, that is, half-sample symmetric padding at both ends of the window. Because the 16-tap filters of the db8 basis extend beyond the window near its edges, the signal must be extended there; mirroring keeps the extended signal continuous at the edge, whereas zero padding would create an artificial step and periodic extension would join the two ends of the window, either of which would produce spurious detail coefficients at the edges. Because each window is also taken from the interior of a recording (Section 2.8.1), the padded boundaries do not coincide with the transients at the start and end of an acquisition. The branch uses features from level-4 and level-5 detail coefficients. These are processed by three parallel convolutions with kernel sizes 3, 5 and 7 (42, 42 and 44 channels, concatenated to 128), three residual blocks and a 1×1 convolution with batch normalization; adaptive average pooling then aligns the feature length with the 3000 input samples, and squeeze-and-excitation (SE) channel weighting, with a bottleneck of 8 channels, follows:

$$z = 1/T \sum_t f_t, \quad w = \sigma(W_2 \text{ReLU}(W_1 z)), \quad \tilde{f}_t = w \odot f_t \quad (3)$$

The resulting channel weights rescale the wavelet features before fusion. DWT-based representations have been used with deep encoders in adjacent ECG tasks [23], and wavelet analysis has an established role in boundary delineation [6]. The basis and levels were fixed a priori and were not tuned. At 500 Hz, the level-4 and level-5 detail coefficients span approximately 15.6–31.3 Hz and 7.8–15.6 Hz, which correspond approximately to the scales used for QRS and T-wave delineation in the wavelet-based delineator of Martínez et al. [6]. In a post hoc sensitivity analysis over three wavelet families (db4, db8 and sym8) and three level pairs (3–4, 4–5 and 5–6) on the out-of-fold LUDB recordings (seed 123), accuracy varies by 1.74 percentage points across the nine configurations; the configuration used is close to, but not at, the best (sym8, levels 5–6; accuracy 94.31% versus 94.01%), and it was retained because the difference comes from a single seed and is of the same order as the seed-to-seed variation of the full model.

2.5. Temporal Modeling and Self-Attention Block

The fused features first pass through a bidirectional LSTM stage [24],

$$\vec{\mathbf{s}}_t = \text{LSTM}_f(\mathbf{u}_t, \vec{\mathbf{s}}_{t-1}), \quad \overleftarrow{\mathbf{s}}_t = \text{LSTM}_b(\mathbf{u}_t, \overleftarrow{\mathbf{s}}_{t+1}), \quad \mathbf{s}_t = \begin{bmatrix} \vec{\mathbf{s}}_t \\ \overleftarrow{\mathbf{s}}_t \end{bmatrix}. \quad (4)$$

The first bidirectional LSTM is followed by a transformer-style self-attention block [25]. The block uses four attention heads and a position-wise feed-forward network (FFN), each with a residual connection. Layer normalization is applied to the input of the attention sublayer (pre-LN) and to the output of the feed-forward sublayer after its residual addition (post-LN) [26]; these are the only layer-normalization operations in the model. A

second bidirectional LSTM processes the resulting representation before classification. With LayerNorm denoting layer normalization, the attention and residual operations are:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \\ z' &= z + \text{MHSA}(\text{LayerNorm}(z)), \\ z'' &= \text{LayerNorm}(z' + \text{FFN}(z')) \end{aligned} \quad (5)$$

The first Bi-LSTM receives the 384-dimensional fused features and has two layers with 64 hidden units per direction, giving a 128-dimensional output. The attention block therefore operates at an embedding dimension of 128 with four heads, so that $d_k = d_v = 32$, and the FFN expands the representation from 128 to 256 dimensions and back to 128. The second Bi-LSTM has one layer with 64 hidden units per direction. The classifier is a per-sample linear layer from 128 to three classes, and the complete model has 9.74 M parameters. The component ablation in Section 3.7 compares the full model with variants that remove or alter individual modules. The ablation is repeated with four training seeds, and the full model and the –DWT variant are extended to eight seeds because the four-seed variance estimate proved unstable.

2.6. Frame Classification and QT Derivation

A classifier maps the representation to per-sample class probabilities $P = \{p_{i,c}\}$. QT is derived beat-wise: each predicted QRS onset is paired with the first predicted T-wave offset that follows it,

$$QT_k = \frac{1000}{f_s} \left(t_{\text{T-off}}^k - t_{\text{QRS-on}}^k \right) [\text{ms}] \quad (6)$$

Here f_s is the sampling frequency. QT is calculated from the predicted landmarks using the forward-pairing rule, subject to the validity criteria in Section 2.10. No separate R-peak detector is used in the evaluation. Reference beats are formed from the database annotations by pairing each annotated QRS complex with the first annotated T wave that follows it. During scoring, the midpoint of the reference QRS complex serves only as an anchor that selects which predicted QRS segment corresponds to a given reference beat; both endpoints of each measured interval, the QRS onset and the T-wave offset, are predicted by the model. Because the anchor determines which beats are scored and which predicted segment is paired with each reference beat, the reported interval errors and coverage are reference-anchored and may be optimistic relative to reference-free deployment, which would require an R-peak detector or an equivalent cardiac-cycle segmentation step (Sections 4.3 and 4.4). Exposing these landmarks permits inspection of the resulting interval; in deployment, the accuracy of the complete measurement pipeline would also depend on cycle detection, as it does on post-processing.

2.7. Focal Loss with a Scheduled Focusing Parameter

To address imbalance among waveform classes, we use the alpha-balanced focal loss of Lin et al. [27] with fixed class weights and a linearly scheduled focusing parameter:

$$L = -\sum_{i=1}^N \alpha_{y_i} (1 - p_{i,y_i})^{\gamma(e)} \log p_{i,y_i}, \quad \gamma(e) = \gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \frac{e}{E_{\text{total}}} \quad (7)$$

The class weights $\alpha = (0.25, 0.25, 0.5)$ for baseline, QRS and T wave were set by hand before any experiment reported here, giving the baseline and QRS classes equal weight and the T-wave class twice that weight; they are not inverse class frequencies. The frame-class distribution of the cross-validation pool is 70.26% baseline, 10.39% QRS and 19.35% T wave, for which normalized inverse-frequency weights would be (0.088, 0.594, 0.319). A post

hoc comparison with these weights over four seeds favored the hand-set weights on every reported metric (for example, out-of-fold LUDB QT MAE 18.42 versus 27.08 ms and QTDB QT MAE 35.29 versus 39.77 ms); it is reported as a check, not as the basis on which the weights were chosen. The focusing parameter increases linearly from $\gamma_{\min} = 0$ towards $\gamma_{\max} = 2$ over a nominal budget of $E_{\text{total}} = 200$ epochs. The gradual increase follows a curriculum rationale: training begins with a class-weighted cross-entropy objective without focal modulation while the waveform classes are learned, and the relative emphasis on poorly classified samples, such as those near waveform boundaries, increases as training proceeds. Because early stopping usually ends training before the nominal budget is exhausted, the value reached at the selected checkpoint is lower than the nominal endpoint: for the 0-to-2 schedule, the realized γ averages 0.65 for seed 123 and ranges from 0.20 to 1.35 across folds and seeds. The label “0 $\rightarrow$ 2” therefore denotes the nominal schedule; the realized value is lower. Section 3.8 compares this schedule with cross-entropy, a fixed focusing parameter and alternative schedules. The 0-to-2 schedule is the configuration used in the main experiments; it is not the best-performing schedule in the reported frame-level sensitivity analysis. It was fixed before the reported experiments and is retained, because adopting a better-performing schedule on the basis of fixed-test-set results would constitute test-set selection. The validation loss used for checkpoint selection and early stopping is the same class-weighted focal loss evaluated with the γ of the current epoch, so its change across epochs reflects both model improvement and the increase in γ .

2.8. Datasets

2.8.1. LUDB

LUDB [14] contains 200 ten-second 12-lead recordings sampled at 500 Hz with cardiologist-annotated waveform boundaries. LUDB assigns each recording to one of seven rhythm classes. The intended selection comprised the four sinus-rhythm classes, namely sinus rhythm (143), sinus bradycardia (25), sinus arrhythmia (8) and sinus tachycardia (4), totaling 180 recordings, and excluded atrial fibrillation (15), atrial flutter (3) and irregular sinus rhythm (2). The exclusion was intended because irregular RR intervals make the QT interval vary from beat to beat and its rate correction unreliable, and because fibrillatory or flutter waves can obscure the end of the T wave in atrial fibrillation and flutter. Owing to an error in mapping the selection list to the recording files, the realized cohort differs from the intended one: it includes 18 recordings outside the four sinus-rhythm classes, namely 16 with atrial fibrillation or flutter and 2 with irregular sinus rhythm. All reported results were computed on this realized cohort. The 180 selected recordings were split 143/18/19 into training, validation and test partitions, and one non-test recording was subsequently excluded because its lead-II waveform boundaries could not be parsed, leaving 179. Nineteen recordings formed a fixed test set; the remaining 160 formed the cross-validation pool, with 128 training and 32 validation recordings per fold. One 6-s window on lead II, the central 2–8 s of the recording, was analyzed per recording; the first and last 2 s are not annotated and were excluded (Section 2.2). Results from the fixed test set and from the cross-validation pool are identified separately.

2.8.2. QTDB

QTDB [15] contains 105 fifteen-minute two-channel recordings originally sampled at 250 Hz. After resampling to 500 Hz, 149 consecutive non-overlapping 6-s windows per recording were analyzed, yielding 15,645 windows and 97,858 reference QRS-onset/T-offset pairs. These windows do not exhaust the full 900-s duration. The first channel was used: ECG1 in 57 records, MLII in 12, CM5 in 12, D3 in 9, V4 in 7, V5 in 4, and V1–V2, V2–V3,

V3 and CM2 in one record each. This evaluates transfer to heterogeneous single-lead recordings rather than lead-II-specific performance.

Two QTDB reference standards are distinguished. Manual final annotations (.q1c) provide expert fiducial locations determined with both leads available to the annotator. Restricting interval evaluation to 6-s windows within the selected annotated segments yields 101 recordings, 530 windows and 2417 beats, averaging 23.9 beats and 5.2 windows per recording. These results are reported separately in Section 3.3. Automatic ecgpuwave annotations for signal 0 (.pu0) provide the reference for the analyzed 15,645 windows and 97,858 beats. The automatic reference is channel-specific; the manual reference is not described here as an independently annotated first-channel reference.

Manual frame labeling additionally requires the onset and offset of each wave. The annotation inventory used in this study contains 3623 fully bracketed QRS complexes and 3542 T waves, of which 1412 have an onset marker. Forty-one recordings have every T wave fully bracketed, eight are partially bracketed and 56 lack T-wave onsets. Expert-reference frame metrics therefore use the 41 fully bracketed recordings, comprising 177 windows and 102,495 T-wave frames. Analyses covering all 105 recordings use the automatic reference. The selected subset and its limitations are examined in Section 3.6.

2.8.3. CNUH

The CNUH cohort consists of 33 routine 12-lead recordings from Chonnam National University Hospital, with QRS and T-wave onset/offset annotations supplied by a cardiologist. Each of the 33 recordings was obtained from a different patient, so the recordings are independent at the patient level. One 6-s lead-II window per recording (the central 2–8 s, as for LUDB) was analyzed, comprising 209 beats. These results use expert annotations. The study was approved by the Institutional Review Board of Chonnam National University Hospital (CNUH-2023-167; 16 June 2023). Data were de-identified, and informed consent was waived owing to the retrospective design. The study conformed to the Declaration of Helsinki.

Both external datasets are evaluated without any fine-tuning.

2.9. Baselines

The comparisons include a rule-based delineator, neural delineation configurations, two U-Net delineators and a scalar regressor. NeuroKit2 [7] supplies wavelet-based QRS onsets and T-wave offsets. Control (– Attention) removes the attention block from QTANet while retaining the residual encoder, DWT branch and both Bi-LSTM stages; it is the same configuration as the – Attention ablation. A further ResNet + LSTM + DWT configuration is reported as an additional delineation comparator; component-level interpretations in this paper rely on the explicitly identified QTANet-versus-control comparison. The scalar-regression baseline follows the QTNet-style formulation [11] and uses the residual convolutional trunk with LSTM and DWT branches and a regression head producing one QT value per 6-s window. It uses the same 160-recording cross-validation pool and fold split, batch size 8, L1 loss on z-score-normalized targets and early-stopping patience of 25 epochs. The raw-target version achieved LUDB MAE 33.72 ms and $r = 0.136$, compared with 33.89 ms for a constant predictor; target normalization improved these to 23.61 ms and $r = 0.618$. Section 3.2 uses the normalized version. These checks demonstrate learning under the selected configuration but do not establish that all regression architectures or training choices have been optimized. QTcNet [13] is discussed as related work rather than reimplemented because its 12-lead input and QTc target differ from the present task.

The U-Net comparators are encoder–decoder delineators trained under the same protocol as QTANet. The first reimplements the one-dimensional U-Net of Moskalenko

et al. [17] to its published specification: four encoder blocks of two convolutions each, with 4, 8, 16 and 32 channels; a 64-channel bottleneck; a symmetric decoder with skip connections joined by zero padding rather than cropping; convolutions with kernel size 9 and padding 4; transposed convolutions with kernel size 8, stride 2 and padding 3; batch normalization and ReLU throughout; and a final 1×1 convolution. The only adaptation is the number of output classes, three rather than four, because the frame labels used here do not include the P wave. This model has 0.167 M parameters, against 9.74 M for QTANet. The second is a larger modern U-Net variant (4,095,555 parameters, 4.10 M) with four encoder stages of 32, 64, 128 and 256 channels, each followed by max pooling with width 2 and stride 2, and a 256-channel bottleneck. Each of the four decoder stages upsamples with a transposed convolution (kernel size 2, stride 2), concatenates the features of the corresponding encoder stage and applies a convolution block, giving 256, 128, 64 and 32 channels. Every convolution block consists of two convolutions with kernel size 7 and padding 3, each followed by batch normalization and ReLU, with dropout (0.2) after the first, and a 1×1 convolution shortcut added to its output; a final 1×1 convolution maps the 32 channels to the three classes. Batch normalization is the only normalization, and the model contains no attention layer. Both were trained on the same folds with the same four seeds, loss, early-stopping rule and post-processing as QTANet, scored with the same M2 matching rule and compared using the same recording-level paired tests (Section 3.11).

2.10. Evaluation Metrics and QT Matching Protocol

Three metric levels are distinguished: sample-wise classification accuracy, precision, recall, F1 and T-wave AUC; tolerance-based waveform detection sensitivity (Se), positive predictive value (PPV), F1 and signed boundary errors; and interval-level QT MAE, correlation, bias where available, and measurement coverage. Frame metrics and tolerance-based detection metrics quantify different properties and are not interchangeable with interval accuracy.

For the primary beat-level interval analyses, the M2 rule pairs each predicted QRS onset with the first subsequent predicted T-wave offset as described in Section 2.6. This candidate-selection step does not select an offset by proximity to a reference offset. M1 instead selects the predicted offset closest to the reference and is used only in sensitivity analyses. Reference-based scoring remains distinct from predicted-landmark pairing. M1 and M2 results are identified separately because their rankings can differ.

A paired interval is retained only if it falls within 50–700 ms. This range was fixed a priori as a plausibility filter for paired landmarks rather than as a physiological limit. Narrower acceptance windows discard large errors and lower the reported MAE; the range is stated explicitly, and the sensitivity of the results to its upper bound is reported in Section 3.3.

Detection rate is the percentage of reference beats with a valid interval under M2 and the stated range. It accompanies the primary beat-level MAE (Section 3.3) because error among successful measurements does not describe the frequency of missing measurements.

For wave detection under the 150 ms tolerance, each reference fiducial (QRS onset, QRS offset, T-wave offset) was matched to the nearest unmatched predicted fiducial of the same type within a ± 150 ms window, processing references in temporal order; a predicted fiducial can be assigned to at most one reference. Matched references count as true positives, unmatched references as false negatives and unmatched predictions as false positives, so that $Se = TP / (TP + FN)$ and $PPV = TP / (TP + FP)$. Boundary errors are signed differences (predicted – reference) over matched pairs. Predicted class sequences were post-processed with the single fixed rule used throughout (morphological closing of 5 samples, gap filling up to 15 samples, removal of segments shorter than 5 samples); no

post-processing parameter was adjusted after inspecting detection results. At 500 Hz, these settings correspond to 10, 30 and 10 ms, well below the duration of a QRS complex or a T wave, and the rule was applied identically to every model and ablation variant. This protocol was fixed before the results of Section 3.1 were produced.

For the scalar-regression comparison, QTANet's available beat-derived QT estimates are averaged within each 6-s window. Both formulations use identical window reference values and are evaluated on the common set of (fold, window) pairs. Section 3.2 reports the mean and SD of the fold-specific window MAEs and correlations over pooled windows, and separately summarizes paired recording-level differences after averaging fold-specific MAEs within each recording (Section 2.11). This comparison is conditional on common valid windows and is distinct from both the beat-level errors and the 101-recording expert-reference QTDB interval subset in Section 3.3.

The recording is the unit used for paired statistical analysis. Beat-, window- and recording-level summaries are identified separately to avoid treating correlated beats as independent observations.

2.11. Statistical Analysis

Two-sided Wilcoxon signed-rank tests are used for the paired per-recording comparisons identified in the tables; exact ties contribute no signed rank. Reported *p*-values are unadjusted, and findings are interpreted as exploratory in view of the multiple endpoints and reference-standard comparisons, using a nominal significance level of 0.05. For the boundary-versus-regression comparison (Section 3.2), fold-specific MAEs are first averaged within each recording for each method, and the recording-level difference is defined as regression minus QTANet. Recordings are then sampled with replacement in 10,000 bootstrap replicates using random seed 0. The 2.5th and 97.5th percentiles of the bootstrap mean differences form the 95% confidence interval. These intervals are distinct from the across-fold SD of the window MAE. For the pooled window-level correlation on QTDB, where each recording contributes many windows, a 95% confidence interval is obtained with a cluster bootstrap that resamples recordings, and within-recording association is quantified with the repeated-measures correlation [28]. Standardized paired effect sizes are not reported.

Training-seed variation is reported separately from recording-level variation. The main repeated analyses use seeds 123, 7, 41 and 42 with the same five-fold split. The multi-seed component ablation, loss-schedule, class-weight and U-Net comparisons use the same four seeds; the eight-seed extension of the ablation adds seeds 1, 2, 3 and 4, and the learning curve uses seeds 123 and 7. Table notes identify analyses that instead use three seeds (123, 41 and 42), as well as single-seed analyses. For repeated-seed paired tests, each recording's metric is averaged over the included seeds before testing. These averages describe the evaluated runs and do not constitute prediction-level ensembling or establish the performance of an arbitrary checkpoint. Models trained on overlapping training folds are not independent training replications; external and fixed-test recordings are also evaluated repeatedly by the fold models. The descriptive agreement analysis in Section 3.3 uses pooled fold-beat pairs. For external data, a reference beat can contribute one prediction from each fold model, so the pair count is not the number of independent beats. The displayed bias and 95% limits of agreement describe the pooled prediction errors; they are not confidence intervals for a population parameter. Section 3.8 reports a descriptive correlation across schedules and folds. Because those observations share fold partitions and evaluation recordings, no independent-sample *p*-value is assigned to that correlation.

2.12. Implementation and Internal Evaluation Protocols

Models were implemented in Python (v3.11.16) with PyTorch (v2.5.1; CUDA 12.4) and trained in FP32 on a single NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) using Adam (learning rate 1×10^{-3} , weight decay 1×10^{-4} , beta = (0.9, 0.999)) and a one-cycle learning-rate schedule with maximum 1×10^{-3} and cosine annealing, stepped per mini-batch. Batch size was 8 and the maximum budget was 200 epochs. For delineation models, early stopping used patience 10 and the checkpoint with the lowest validation loss; training typically ended after 60–100 epochs. The class weights in Equation (7) were (0.25, 0.25, 0.5) for baseline, QRS and T. The focusing parameter increased from 0 towards 2 over the planned 200 epochs, so early-stopped runs did not reach the planned endpoint (Section 2.7). The scalar regressor instead used the loss and patience described in Section 2.9. The DWT branch used db8 levels 4–5. The first Bi-LSTM had two layers and the second had one, each with 64 hidden units per direction. Attention used four heads and dropout 0.1 on the attention weights. Dropout was 0.2 in the residual convolutional blocks and 0.3 between the two layers of the first Bi-LSTM, in the feed-forward sublayer and before the classifier; the DWT branch and the second Bi-LSTM had no dropout. Inputs were single-lead 6-s windows at 500 Hz. The shuffled fold split used seed 42; seed 123 was used for the single-seed results, with repeated runs identified in the tables. All compared configurations used the same fold partitions. All settings were fixed before the reported experiments, and none was selected using the fixed test set or either external cohort. The encoder width, recurrent and attention dimensions, optimizer settings, early-stopping patience, post-processing rule and matching tolerance were fixed a priori; the wavelet settings, class weights, focal schedule and validity window were also fixed a priori and were examined only in post hoc sensitivity analyses.

Internal evaluation comprised a separate 19-recording fixed test set and a 160-recording non-nested out-of-fold (OOF) analysis covering 1001 beats. The fixed test set was not used for training or early stopping. For OOF evaluation, each recording was excluded from gradient-based training, but its held-out fold also supplied the early-stopping loss. A nested sensitivity analysis used separate optimization, early-stopping and evaluation subsets of 112, 16 and 32 recordings. Its QTANet frame metrics were $93.10 \pm 1.11\%$ accuracy, $90.82 \pm 1.35\%$ macro F1 and 0.9865 T-wave AUC, compared with $93.20 \pm 0.53\%$, $91.07 \pm 0.86\%$ and 0.9894 for the standard protocol on the separate fixed test set (Section 3.1). Corresponding QTANet-minus-control margins were +0.47 versus +0.18 accuracy points, +0.87 versus +0.61 macro-F1 points and +0.0071 versus +0.0041 T-wave AUC. Because the two protocols were evaluated on different recordings, these comparisons are descriptive rather than direct protocol contrasts. This sensitivity analysis concerns frame metrics; fixed-test and non-nested OOF interval results are reported separately.

3. Results

3.1. Waveform Boundary Detection

Table 1 summarizes QTANet's frame-level classification as means and standard deviations across the five per-fold models. Figure 3 presents ROC curves and row-normalized confusion matrices for LUDB out-of-fold (OOF) predictions and the two external datasets. The LUDB panel in Figure 3 uses the 160-recording OOF pool, whereas Table 1 uses the separate 19-recording fixed test set. Figure-level AUC values and across-fold mean AUC values are identified separately. QTDB uses the automatic reference for this comparison. The recording-level comparison, including the expert-annotated QTDB subset, is reported separately in Section 3.6.

Table 1. Frame-level waveform classification of QTANet (five-fold, mean $\pm$ SD, seed 123).

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	T-Wave AUC
LUDB (19 recordings)	93.2 $\pm$ 0.5	90.9 $\pm$ 1.0	91.4 $\pm$ 1.5	91.1 $\pm$ 0.9	0.989 $\pm$ 0.004
QTDB (105 recordings)	86.5 $\pm$ 0.7	85.1 $\pm$ 0.6	84.1 $\pm$ 1.4	84.5 $\pm$ 0.8	0.954 $\pm$ 0.004
CNUH (33 recordings)	88.8 $\pm$ 0.9	87.4 $\pm$ 0.9	85.6 $\pm$ 1.8	85.9 $\pm$ 1.3	0.973 $\pm$ 0.003

Evaluation scale: LUDB fixed test set, 19 recordings and 120 beats; QTDB automatic reference, 105 recordings, 15,645 windows and 97,858 beats; CNUH expert reference, 33 recordings and 209 beats. The reported four-seed out-of-fold LUDB summary is $93.59 \pm 1.08\%$ accuracy, $91.67 \pm 1.27\%$ macro F1 and 0.990 ± 0.004 T-wave AUC. These out-of-fold models used the held-out fold for early stopping (Section 2.12). Across seeds 123, 7, 41 and 42, reported LUDB out-of-fold accuracy spans 93.21–94.01% and macro F1 91.18–91.88%; QTDB accuracy spans 83.99–86.55% and macro F1 80.74–84.29%; CNUH accuracy spans 87.66–89.07%. Frame-level three-class F1 is distinct from the tolerance-based detection F1 in Table 2.



Figure 3. Frame-level waveform classification for QTANet and Control (–Attention). (a,d,g) ROC curves for QRS and T wave on LUDB (OOF), QTDB and CNUH, respectively; (b,c,e,f,h,i) the corresponding row-normalized confusion matrices. Solid and dashed curves denote QRS and T wave, respectively; colors distinguish models. Matrix entries are percentages within each reference class and may not sum to exactly 100% because of rounding. LUDB and CNUH use expert annotations; QTDB uses the automatic reference. The LUDB OOF panel is distinct from the fixed-test summary in Table 1.

Table 2. Wave detection under a 150 ms tolerance (seed 123). LUDB uses out-of-fold predictions from 160 recordings; QTDB uses the automatic full-coverage reference; CNUH uses expert annotations.

Dataset	Model	QRS Se/PPV/F1 (%)	T-Offset Se/PPV/F1 (%)	QRS-Onset Error (ms)	T-Offset Error (ms)
LUDB (public, 200) †	U-Net, 12-lead score averaging [17]	100.00/99.93/99.96	99.77/99.46/99.61	+1.5 ± 11.1	−2.4 ± 30.4
LUDB (OOF, 160)	QTANet	99.50/95.61/97.52	99.21/95.23/97.18	+1.02 ± 11.08	+7.07 ± 24.44
LUDB (OOF, 160)	Control (−Attention)	99.80/92.63/96.08	98.31/89.84/93.88	+2.90 ± 10.43	+4.23 ± 21.29
LUDB (OOF, 160)	ResNet + LSTM + DWT	99.70/91.77/95.57	97.62/89.69/93.49	+1.15 ± 10.24	+4.15 ± 23.51
QTDB (105)	QTANet	94.86/95.51/95.18	90.99/85.44/88.13	+6.11 ± 17.85	−0.03 ± 39.35
QTDB (105)	Control (−Attention)	96.28/82.73/88.99	82.20/72.36/76.97	+5.50 ± 19.01	−6.23 ± 41.19
QTDB (105)	ResNet + LSTM + DWT	97.16/83.51/89.82	84.79/75.58/79.92	+4.18 ± 20.00	−9.15 ± 42.07
CNUH (33)	QTANet	95.89/98.44/97.15	91.26/95.03/93.11	+3.41 ± 13.89	+8.39 ± 28.76
CNUH (33)	Control (−Attention)	98.93/95.52/97.19	94.32/89.18/91.68	+5.10 ± 12.01	+6.61 ± 33.37
CNUH (33)	ResNet + LSTM + DWT	99.20/97.12/98.14	93.78/89.36/91.52	+2.60 ± 11.41	+3.48 ± 29.95

Matching follows Section 2.10. LUDB uses 160 out-of-fold recordings and 1001 beats; checkpoint-selection exposure is described in Section 2.12. Boundary errors are signed predicted-minus-reference differences among matched fiducials. The QRS detection column refers to the reported QRS landmark evaluation. † Moskalenko et al. [17]: U-Net applied to individual leads with score averaging over 12 leads, trained on 255 additional non-public recordings and evaluated on public LUDB. F1 is calculated from the published Se and PPV. The source also reports lead-II-only results. Differences in training data, evaluation records, lead aggregation and processing preclude a direct ranking or attribution of PPV differences solely to lead count.

On QTDB, attention increases QRS PPV from 82.73% to 95.51% and T-offset PPV from 72.36% to 85.44%. QRS sensitivity decreases from 96.28% to 94.86%, whereas T-offset sensitivity increases from 82.20% to 90.99%. Thus, the attention configuration shows both higher precision and improved T-offset sensitivity against the automatic reference. Smaller PPV gains occur in the internal data and in the clinical cohort. The corresponding paired recording-level comparison is reported in Table 3. These results are consistent with fewer false-positive waveform predictions, and Section 3.10 analyzes the number, duration and location of the spurious segments directly.

Table 3. Recording-level paired comparison of wave detection, QTANet versus Control (−Attention). QTDB uses the automatic reference; LUDB and CNUH use expert annotations. *p*-values are unadjusted.

Dataset	Landmark	Metric	Δ (Percentage Points)	Wilcoxon <i>p</i>	Win/Tie/Loss
QTDB	T-wave offset	F1	+11.55	2.5×10^{-15}	95/0/10
QTDB	T-wave offset	PPV	+12.63	9.4×10^{-15}	—
QTDB	QRS	F1	+4.96	1.7×10^{-11}	81/0/24
LUDB (OOF)	T-wave offset	F1	+2.67	3.6×10^{-4}	43/103/14
LUDB (OOF)	QRS	F1	+1.35	3.4×10^{-4}	25/130/5

Table 3. Cont.

Dataset	Landmark	Metric	Δ (Percentage Points)	Wilcoxon p	Win/Tie/Loss
CNUH	T-wave offset	F1	+1.39	5.4×10^{-3}	20/7/6
CNUH	QRS	F1	−0.09	0.49	6/16/11

Δ is defined as QTANet minus Control (−Attention), in percentage points. Win/tie/loss counts recordings in which QTANet has a higher/equal/lower metric value than Control. Differences are means of per-recording differences, so they need not equal differences between the pooled values in Table 2. Ties indicate equal metric values, including recordings with perfect detection by both models. Out-of-fold LUDB results are subject to the checkpoint-selection qualification in Section 2.12.

Figure 4 illustrates QRS and T-wave labeling in a representative LUDB segment. This qualitative example is not an estimate of population-level accuracy.

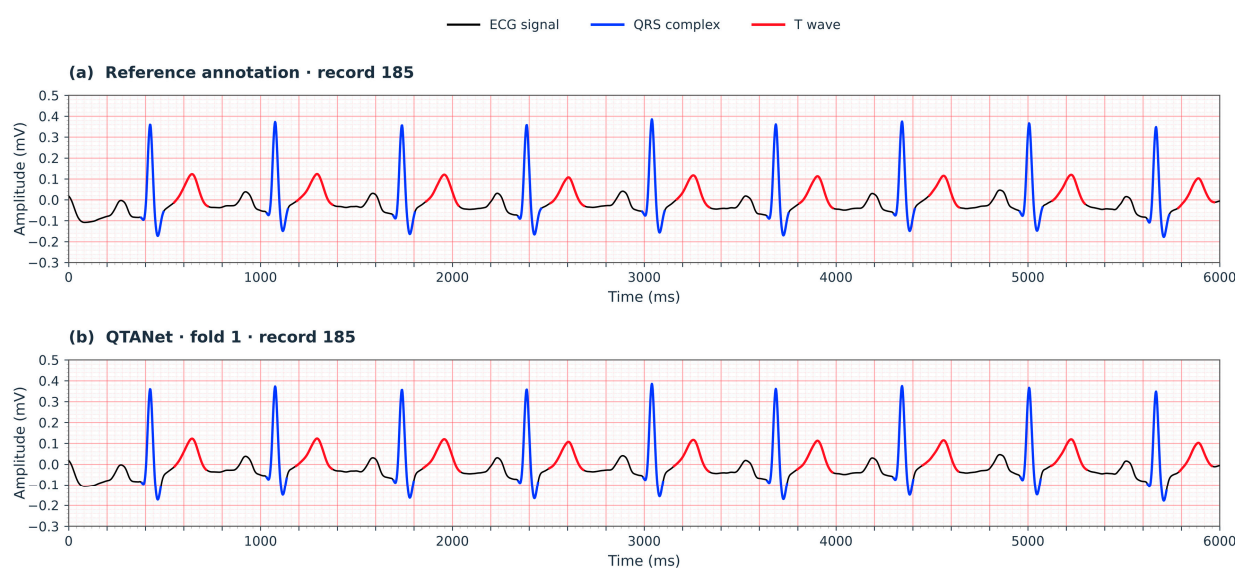


Figure 4. Representative delineation example (LUDB, recording 185, lead II, 6-s segment). (a) Expert-annotated QRS (blue) and T-wave (red) segments; (b) model prediction. QT is derived from the QRS onset and T-wave offset implied by the predicted segmentation.

3.2. Boundary-Derived QT Versus Scalar Regression

Table 4 and Figure 5 compare boundary-derived QT with scalar regression using identical window references and common (fold, window) pairs. QTDB reference boundaries were obtained from the automatic (.pu0) annotations corresponding to the first signal channel. Record sel33 had 454 paired reference QT intervals across 149 windows, all between 720 and 888 ms. None satisfied the reference validity criterion ($50 < QT < 700$ ms), so no valid window-level reference QT mean could be formed. Record sel33 was therefore excluded, leaving 104 recordings. This exclusion was based on reference validity, not prediction failure or observed model error. The long reference intervals of sel33 are genuine: its cardiologist (.q1c) annotations give a median QT of 764 ms, with every annotated beat at or above 700 ms. More generally, 3328 automatic-reference beats (3.40%) from 41 recordings exceed 700 ms. Comparison with the manual annotations shows that these comprise both truly prolonged intervals (sel33 and sele0116, with cardiologist median QT of 764 and 726 ms) and automatic-annotation errors of 200 ms or more (sel47, sel50 and sel14172, with cardiologist medians of 410–488 ms against automatic medians of 614–752 ms). The validity criterion therefore removes a mixture of truly long-QT recordings and annotation artifacts, not artifacts alone; its effect on the reported error is quantified in Section 3.3. Full frame-classification evaluation retains all 105 recordings. The expert-reference interval evaluation in Section 3.3 is a separate analysis.

Table 4. Window-level QT MAE (ms), mean $\pm$ SD across five folds, for boundary estimation and scalar regression on common (fold, window) pairs with identical window reference values. LUDB uses non-nested OOF predictions; external models are evaluated without fine-tuning. QTANet values are pooled over the four training seeds (123, 7, 41 and 42) without averaging predictions across seeds; the regression baseline is a single seed-123 run.

Dataset	Method	QT MAE (ms), Mean $\pm$ SD
LUDB (out-of-fold)	QTANet (boundary)	13.19 $\pm$ 3.16
LUDB (out-of-fold)	Regression (scalar)	23.61 $\pm$ 5.83
QTDB (external)	QTANet (boundary)	32.73 $\pm$ 2.35
QTDB (external)	Regression (scalar)	56.77 $\pm$ 8.31
CNUH (clinical)	QTANet (boundary)	19.40 $\pm$ 1.69
CNUH (clinical)	Regression (scalar)	39.80 $\pm$ 5.20

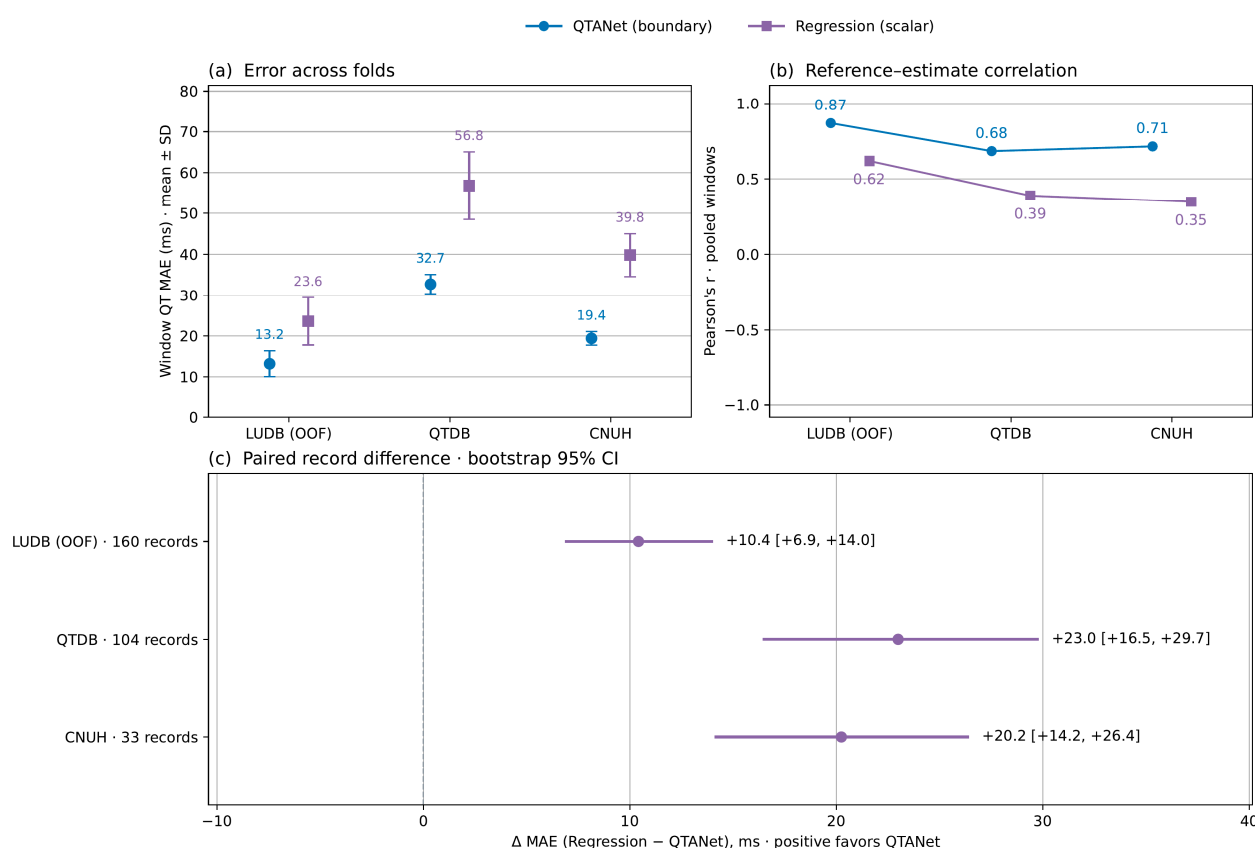


Figure 5. Boundary-derived QT estimation versus scalar regression using identical window references and common (fold, window) pairs. (a) Window QT MAE, mean $\pm$ SD across five folds (Table 4). (b) Pearson correlation over pooled windows, which mainly reflects between-recording differences in mean QT; within-recording association on QTDB is reported in Section 3.2. (c) Mean paired recording-level difference in window QT MAE (regression minus QTANet), with percentile bootstrap 95% confidence intervals based on 10,000 recording-level resamples (seed 0; Table 5). QTANet values are averaged over the four training seeds (123, 7, 41 and 42), with absolute errors rather than predictions averaged across seeds for each recording; the regression baseline is a single seed-123 run. Both methods are scored on the same (fold, window) pairs with identical window reference values, using matching rule M2 with the 50–700 ms window applied to reference and predicted intervals. The paired analysis comprises 160 LUDB OOF, 104 QTDB and 33 CNUH recordings. The panel-specific aggregation units differ; QTDB uses automatic (.pu0) annotations and excludes sel33 because no reference QT interval satisfies $50 < QT < 700$ ms (Section 3.2); it is separate from the expert-reference interval subset in Section 3.3.

Table 5. Paired recording-level difference in window QT MAE on common (fold, window) pairs. Differences are regression minus QTANet, so positive values favor QTANet. Bootstrap 95% confidence intervals use 10,000 recording-level resamples (seed 0); values are reported to the precision shown in Figure 5c. QTANet errors are averaged over the four training seeds before pairing with the single seed-123 regression run, and the 50–700 ms window is applied to reference and predicted intervals in all three cohorts.

Dataset	Recordings (<i>n</i>)	Mean Δ MAE (ms)	95% CI Lower (ms)	95% CI Upper (ms)
LUDB (OOF)	160	+10.4	+6.9	+14.0
QTDB	104	+23.0	+16.5	+29.7
CNUH	33	+20.2	+14.2	+26.4

Each recording contributes one difference after its fold-specific MAEs are averaged within each method. The equal-recording weighting can yield a different difference from the across-fold means in Table 4. The analysis includes 160 LUDB OOF, 104 QTDB and 33 CNUH recordings.

Delineation estimates average the available beat-derived QT values within each window. MAE is calculated on common windows separately for each fold and is then summarized across folds. These values are distinct from the beat-level errors in Section 3.3 and from the recording-level differences in Table 5. The comparison is conditional on valid estimates from both methods; it does not quantify performance on windows without a common valid prediction. For seed 123, common (fold, window) pairs comprise 100.0% of the evaluated pairs for LUDB OOF (160 of 160), 95.6% for QTDB (74,753 of 78,225) and 100.0% for CNUH (165 of 165); the remainder lack a valid QTANet estimate or a valid reference QT (as for sel33), whereas the regression model returns an estimate for every window. A comparison restricted to commonly measured windows can favor every method evaluated on it (Section 3.5).

The boundary model has lower mean window QT MAE than the scalar baseline on all three datasets: 13.19 versus 23.61 ms on LUDB OOF, 32.73 versus 56.77 ms on QTDB, and 19.40 versus 39.80 ms on CNUH. Figure 5b also shows higher pooled-window correlation for boundary estimation. Both methods have larger errors on the external datasets than on LUDB. These results support lower error for the boundary formulation under the evaluated training budget and common-window protocol; they do not establish a general failure of scalar regression or isolate a causal explanation for the transfer difference.

The pooled correlations in Figure 5b, however, mainly reflect between-recording differences in mean QT. On QTDB, the only dataset with multiple windows per recording, the pooled window-level *r* of QTANet for seed 123 is 0.736 (cluster-bootstrap 95% CI 0.648–0.811), whereas the repeated-measures correlation is 0.162 and the mean within-recording *r* is 0.252. Within-recording tracking is therefore weaker than the pooled correlation suggests (Section 4.4). LUDB and CNUH contribute one window per recording, so within-recording correlations cannot be estimated for them; the five fold-model predictions of the same window reflect model variance rather than within-recording variation.

Table 5 and Figure 5c summarize the paired recording-level comparison. Mean differences in window QT MAE (regression minus QTANet) are +10.4 ms for LUDB OOF, +23.0 ms for QTDB and +20.2 ms for CNUH. The bootstrap 95% confidence intervals are [6.9, 14.0], [16.5, 29.7] and [14.2, 26.4] ms, respectively. All three intervals exclude zero in the direction favoring QTANet. The bootstrap confidence intervals quantify recording-level uncertainty conditional on the evaluated trained models; they do not capture the unmeasured training-seed variability of the regression baseline. QTANet errors are averaged over the four training seeds (123, 7, 41 and 42) before pairing with the single seed-123 regression run; for each recording, the absolute errors are averaged across seeds, not the predictions. The evaluated (fold, window) pairs are the same as for seed 123 (160 LUDB

OOB, 74,753 QTDB and 165 CNUH windows) and do not depend on the number of seeds, and the 50–700 ms window is applied to reference and predicted intervals.

Dense waveform labeling and scalar regression use the available annotations differently: one provides a class at each sample, while the other supplies one target per window. Over the cross-validation pool, frame labeling yields 3000 targets per 6-s window, approximately 4.8×10^5 in total, whereas interval regression yields 160. This difference in supervision density may contribute to the transfer gap, but the experiment does not isolate it from differences in architecture and optimization. This comparison concerns how the implemented models use the available expert-annotation dataset, not equal annotation cost or general superiority at all training scales. QTNet uses more than three million device-labeled ECGs [12], and QTcNet uses approximately 1.2×10^5 algorithm-labeled ECGs [13]. A learning curve for the boundary formulation, trained on 25%, 50%, 75% and 100% of each fold's training recordings with two seeds, shows that its external error has not saturated at this training-set size: QTDB QT MAE decreases from 58.50 ms with 32 training recordings to 32.87 ms with 128, whereas internal gains diminish (LUDB QT MAE improves by 4.2 ms from 50% to 75% of the training recordings and by 1.4 ms from 75% to 100%). The conclusion is confined to a limited expert-annotation budget; a regression model trained at the scale of QTNet or QTcNet may not show the transfer difference observed here. Additional regression configurations and a regression learning curve would be required to separate supervision density from architecture, optimization and training-set size.

3.3. QT Accuracy and Measurement Coverage

Table 6 reports QT estimation for all four methods under the M2 matching rule and the 50–700 ms validity range, together with the detection rate. QTDB appears twice, once against each reference standard (Section 2.8).

Table 6. Beat-level QT estimation under M2 and the 50–700 ms validity range. Where shown, mean $\pm$ SD summarizes the five per-fold models. LUDB uses the fixed 19-recording test set.

Dataset (Reference)	Method	QRS-Onset MAE (ms)	T-Offset MAE (ms)	QT MAE (ms)	r	Detection (%)
LUDB (expert)	QTANet	4.50 $\pm$ 0.06	23.08 $\pm$ 4.12	23.37 $\pm$ 4.23	0.631	96.8
LUDB (expert)	Control (– Attention)	4.78 $\pm$ 0.62	17.96 $\pm$ 5.48	17.76 $\pm$ 5.44	0.572	95.5
LUDB (expert)	ResNet + LSTM + DWT	4.26 $\pm$ 0.24	16.00 $\pm$ 2.73	16.53 $\pm$ 2.59	0.749	92.8
LUDB (expert)	NeuroKit2 (dwt)	10.84	32.57	39.41	0.376	90.8
QTDB (expert, .q1c)	QTANet	11.16	32.30	34.94	0.616	95.2
QTDB (expert, .q1c)	Control (– Attention)	11.35	37.81	40.55	0.455	88.2
QTDB (expert, .q1c)	ResNet + LSTM + DWT	12.32	37.68	40.49	0.465	90.5
QTDB (expert, .q1c)	NeuroKit2 (dwt)	32.75	44.11	52.41	0.362	68.7
QTDB (automatic, analyzed windows)	QTANet	11.35 $\pm$ 0.59	27.29 $\pm$ 1.17	31.58 $\pm$ 1.43	0.690	92.8
QTDB (automatic, analyzed windows)	Control (– Attention)	11.81 $\pm$ 1.29	33.22 $\pm$ 4.49	39.05 $\pm$ 4.28	0.557	84.1
QTDB (automatic, analyzed windows)	ResNet + LSTM + DWT	11.79 $\pm$ 1.17	33.59 $\pm$ 7.31	39.20 $\pm$ 6.21	0.572	86.7
QTDB (automatic, analyzed windows)	NeuroKit2 (dwt)	25.03	41.20	49.65	0.391	68.2
CNUH (expert)	QTANet	9.60 $\pm$ 0.61	24.51 $\pm$ 3.31	23.67 $\pm$ 3.12	0.633	97.1

Table 6. Cont.

Dataset (Reference)	Method	QRS-Onset MAE (ms)	T-Offset MAE (ms)	QT MAE (ms)	<i>r</i>	Detection (%)
CNUH (expert)	Control (– Attention)	9.72 ± 0.50	30.31 ± 7.12	30.40 ± 7.36	0.520	98.7
CNUH (expert)	ResNet + LSTM + DWT	9.13 ± 0.46	28.45 ± 4.35	29.26 ± 4.04	0.554	99.8
CNUH (expert)	NeuroKit2 (dwt)	20.91	42.68	52.39	0.349	81.8

The expert QTDB block contains 101 recordings, 530 windows and 2417 beats; the automatic block contains 105 recordings, 15,645 analyzed windows and 97,858 beats. Detection is the percentage of reference beats with a valid reported interval. MAE is therefore conditional on successful measurement. The manual-reference block reports point estimates; the corresponding same-window comparison is described below. Figure 6 provides a separate descriptive agreement analysis using LUDB OOF and external pooled fold-beat pairs. Its QTDB panel is distinct from the 101-recording expert-reference block in this table.

Within the expert-annotated QTDB interval subset, QTANet has the lowest reported QRS-onset MAE (11.16 ms), T-offset MAE (32.30 ms) and QT MAE (34.94 ms), the highest correlation (0.616), and the highest detection rate (95.2%) among the four tabulated methods. Its QT MAE is 5.6 ms lower than the deep comparators and 17.5 ms lower than NeuroKit2, with 26.5 percentage points higher coverage than NeuroKit2. A sensitivity analysis using M1 gives QT MAEs of 29.53, 31.20, 33.18 and 37.38 ms for QTANet, Control, ResNet + LSTM + DWT and NeuroKit2, respectively. This supports the QT-MAE ordering under those two pairing rules; it does not establish that every metric has the same ordering on every cohort.

Scoring the same 530 selected QTDB windows against each reference yields QTANet MAE of 33.31 ± 3.15 ms with automatic annotations and 34.94 ± 2.57 ms with expert annotations. The difference is 1.63 ms. Compared with the automatic-reference value of 31.58 ms across all analyzed windows, the reported difference associated with window selection is 1.73 ms. Thus, both reference standard and window selection contribute to the observed difference. Similar mean errors on selected windows do not establish interchangeability of manual and automatic annotations or beat-wise agreement between them.

The upper bound of the validity range affects the absolute QTDB figures. Against the automatic reference, relaxing the range from 50–700 ms to 50–1000 ms increases coverage from 92.8% to 96.2% and QTANet QT MAE from 31.58 to 39.50 ms (39.53 ms without any restriction). The difference between QTANet and Control is preserved and widens slightly, from 7.47 to 8.50 ms (8.55 ms without restriction), computed, as in Table 6, as means over the five per-fold models. No LUDB or CNUH reference beat exceeds 700 ms, so the upper bound excludes no reference beat in those cohorts; among predicted intervals, it removes a single out-of-fold LUDB beat (1 of 1002), leaving the 1001 beats of the LUDB OOF analysis (Section 2.12). Admitting long intervals changes the signed bias of QTANet, computed for seed 123 over the beats pooled across the five per-fold models, from -7.26 to -15.19 ms. Because the admitted reference intervals include automatic-annotation errors as well as truly prolonged QT (Section 3.2), this shift cannot be attributed to the model alone, but under-measurement of the longest QT intervals cannot be excluded (Section 4.3).

QTANet has the lowest QT MAE among the deep models on both external cohorts in Table 6. On the fixed LUDB test set, however, Control has lower beat-pooled MAE than QTANet (17.76 versus 23.37 ms). Averaging over recordings reduces this gap to 0.5 ms (24.75 ms for Control and 25.24 ms for QTANet; Wilcoxon $p = 0.087$), but the near-equality is produced by a single recording with three annotated beats on which Control fails (187.6 versus 62.6 ms for QTANet) rather than by agreement between the models: the recording-level medians are 10.12 ms for Control and 11.53 ms for QTANet, QTANet

has the lower error in only 6 of 19 recordings, and symmetric trimming of one, two and three recordings from each tail leaves Control ahead by 5.96, 4.76 and 3.18 ms ($p = 0.040$, 0.030 and 0.017). On this test set, Control has the lower QT error. The difference arises at the T-wave offset (beat-pooled error 17.99 ms for Control and 23.05 ms for QTANet), whereas the QRS-onset error is slightly lower for QTANet (4.50 versus 4.78 ms beat-pooled, as in Table 6; 4.57 versus 5.12 ms averaged over recordings). ResNet + LSTM + DWT, trained with a single seed (123), also has lower beat-pooled MAE than QTANet on the fixed test set (16.53 versus 23.37 ms) and keeps this advantage when errors are averaged over recordings (17.79 versus 25.24 ms), although a per-recording paired test does not separate the two models (Wilcoxon $p = 0.16$; QTANet has the lower error in 7 of 19 recordings). No multi-seed estimate is available for this comparator, and its out-of-fold QT error is not reported here. Without post-processing, the QTANet MAE is 22.74 ms. On the out-of-fold set of 160 recordings, averaged over four seeds, QTANet has higher macro F1 than Control (91.69% versus 90.16%) and lower QT MAE (18.42 versus 20.13 ms). For seed 123, the reported out-of-fold values over 160 recordings are similar for QTANet and Control (17.75 and 17.90 ms); the corresponding three-seed recording-level summary is given in Table 7. These estimates retain the early-stopping qualification in Section 2.12. The internal advantage of QTANet is small: its out-of-fold margin over Control (1.71 ms) is about 2.4 times smaller than its four-seed margin on QTDB (4.17 ms; Section 3.7).

Table 7. Recording-level comparison of QT MAE, QTANet versus Control (–Attention). p -values are unadjusted.

Dataset (Reference)	n Rec	QTANet	Control	Δ MAE	Wilcoxon p	QTANet Lower
QTDB (automatic, analyzed windows)	103	40.94	48.82	−7.88	1.1×10^{-11}	83.5%
QTDB (expert, .q1c) †	98	—	—	−7.73	5.4×10^{-4}	65.3%
LUDB (out-of-fold) †	160	18.86	20.45	−1.59	0.024	56.2%
CNUH (expert)	33	25.31	30.55	−5.23	0.014	63.6%

Per-recording MAE is expressed in ms. Results average seeds 123, 7, 41 and 42, except rows marked †, which use seeds 123, 41 and 42. These seed sets differ and should not be treated as identical replications. The paired analysis includes recordings contributing comparable metrics; its denominators differ from the dataset totals. Recording-level means weight each recording equally and may differ from beat-pooled results. Each seed-specific difference reported in the text is computed on the recordings with valid paired measurements for that seed, and these sets are not identical across seeds; the mean of the seed-specific differences therefore need not equal the value obtained by averaging each recording's metric over seeds before testing. Under M1, the expert-reference QTANet-minus-Control difference is +2.14 ms ($p = 5.2 \times 10^{-3}$), illustrating sensitivity to pairing. The M2 result is the primary interval comparison. Differences use unrounded values. The QTDB expert row comprises 98 of the 101 recordings in the expert interval subset of Table 6.

Figure 6 distinguishes two sets of valid measurements. The distribution panels use pairs valid for both QTANet and Control, whereas the Bland–Altman panels use all included QTANet pairs. LUDB contributes OOF predictions; external predictions are pooled across fold models. The displayed QTANet bias and 95% limits of agreement are +5.8 ms [−57.0, +68.6] for LUDB OOF, −7.3 ms [−113.4, +98.9] for QTDB, and +9.5 ms [−68.8, +87.8] for CNUH. These pooled summaries describe agreement among successful measurements. They do not quantify uncertainty due to repeated observations within recordings or establish clinical interchangeability, and they do not replace the error and coverage results in Table 6.

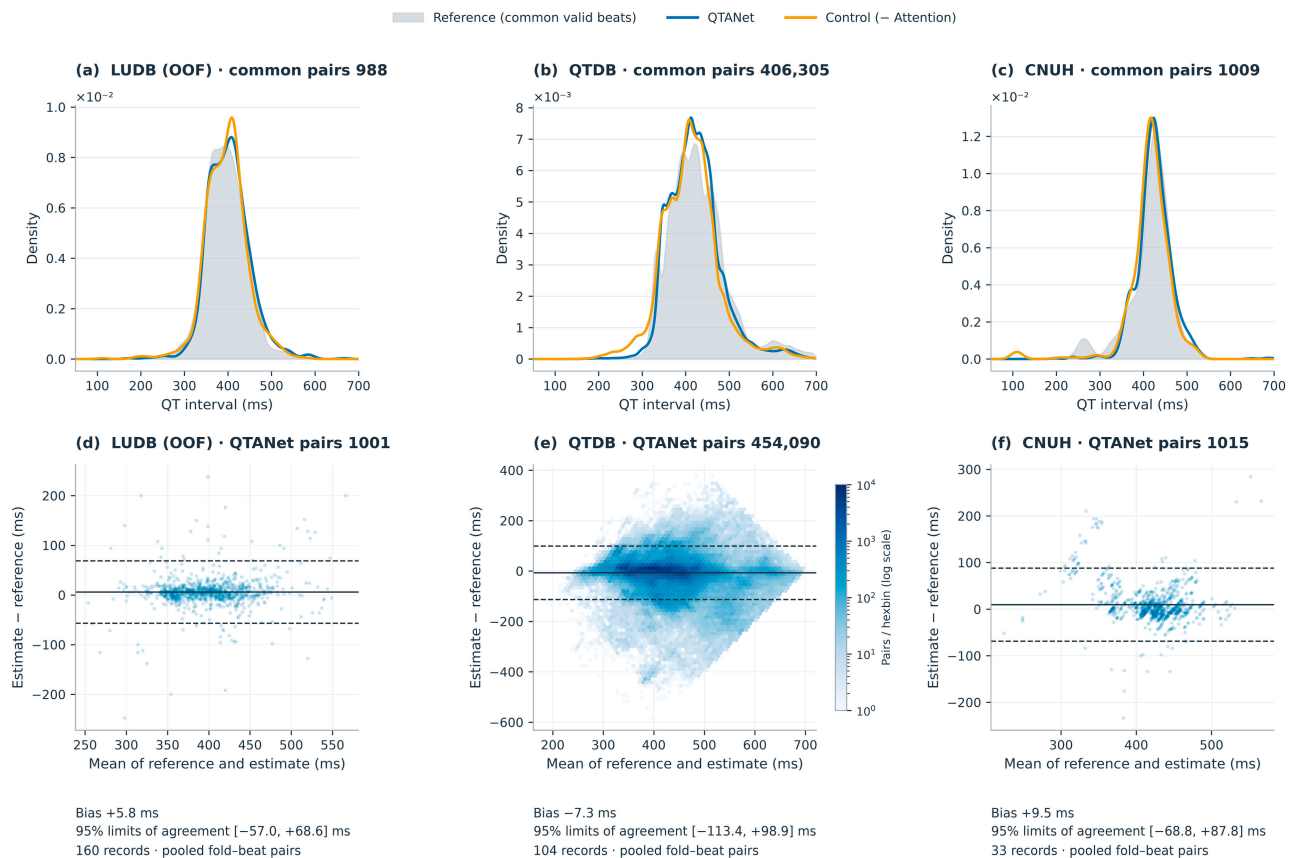


Figure 6. QT distributions and descriptive agreement. (a–c) Reference, QTANet and Control distributions on common valid fold-beat pairs: 988 for LUDB OOB, 406,305 for QTDB and 1009 for CNUH. (d–f) Bland–Altman plots for 1001, 454,090 and 1015 QTANet pairs from 160, 104 and 33 recordings, respectively. Horizontal solid lines denote bias and dashed lines the 95% limits of agreement; differences are estimate minus reference. Panel (e) uses hexagonal bins with a logarithmic count scale. External pair counts include repeated predictions of reference beats by different fold models and are not counts of independent beats. LUDB uses OOB predictions, rather than the fixed test set in Table 6.

3.4. Statistical Comparison at the Record Level

Table 7 compares derived interval errors at the recording level. This is distinct from the frame-label comparison in Section 3.6 because interval estimates also depend on beat pairing and post-processing. The included seed sets and reference standards are identified in the table notes.

Under M2, QTANet has lower per-recording QT MAE on the expert QTDB subset by 7.73 ms, with 65.3% of 98 recordings favoring QTANet. Against the automatic reference, the difference is 7.88 ms, with 83.5% of 103 recordings favoring QTANet. The internal difference is smaller (1.59 ms) and 56.2% of recordings favor QTANet. Statistical evidence for a small mean difference does not establish a clinically important or uniformly reproducible benefit.

On CNUH, the reported per-recording QT MAE difference is 5.23 ms ($p = 0.014$), with 63.6% of recordings favoring QTANet. The T-wave F1 increase is 0.0117 ($p = 0.012$; Section 3.6), the smallest frame-level T-wave gain among the three datasets. Its sign varies across runs: +0.033 for seed 7 and −0.005 for seed 42. This single-center cohort supplies external evidence with limited sample size and seed consistency; the data do not establish that recording cleanliness or a performance ceiling explains the smaller frame-level effect.

On QTDB, the reported Control-minus-QTANet QT MAE differences are +12.68 ms for seed 7 ($p = 1.6 \times 10^{-11}$), +9.86 ms for seed 123 ($p = 1.3 \times 10^{-10}$), +6.58 ms for seed 41 ($p = 6.8 \times 10^{-7}$), and +1.02 ms for seed 42 ($p = 0.39$). All point estimates favor QTANet, but

evidence for an interval-level difference is absent in the fourth run. That run had validation loss 0.0323 and ended at 83 epochs, with per-fold accuracy 0.922–0.943; it was not excluded. Its QTDB T-wave F1 difference remained +0.043. Single-seed tables and repeated-seed summaries are therefore interpreted separately. Prediction-level ensembling is a possible future study, not an evaluated result of averaging per-recording metrics.

3.5. Comparison on NeuroKit2-Detected Beats

Under M2 and the 50–700 ms range, NeuroKit2 returns an interval for 90.8% of LUDB beats, 68.2% of automatic-reference QTDB beats and 81.8% of CNUH beats, compared with 96.8%, 92.8% and 97.1% for QTANet. The corresponding M1 NeuroKit2 coverage values are 85.8%, 62.0% and 71.3%. Table 8 reports the additional analysis restricted to the NeuroKit2-detected subset. This is a conditional error analysis; it does not by itself establish that all models produced valid predictions on an identical beat set in every fold.

Table 8. Reported QT MAE within the NeuroKit2-detected subset (M2; 50–700 ms). QTDB uses the automatic reference. Subset sizes refer to the reported NeuroKit2-detected beats.

Dataset	NeuroKit2 Subset (<i>n</i>)	QTANet	Control (–Attention)	NeuroKit2
LUDB	109	22.90	17.41	39.41
QTDB	66,697	27.15	33.69	49.65
CNUH	171	23.44	25.14	52.39

These results are descriptive. Uncertainty intervals and a model-by-model accounting of valid-beat membership are not reported, so the table is not used as an additional paired significance test or proof that coverage effects have been fully removed.

Within the reported subsets, QTANet has lower MAE than NeuroKit2 by 22.5 ms on QTDB and 29.0 ms on CNUH. Control remains more accurate than QTANet on LUDB. These conditional errors complement the coverage figures, but do not establish the difficulty or clinical risk of beats for which no measurement is returned. Conditioning on commonly measured beats can be substantial: on QTDB, the QTANet error is 27.15 ms on the NeuroKit2-detected subset but 31.58 ms over all beats it measures (Table 6), a difference of 4.4 ms, whereas the corresponding differences are 0.5 ms on LUDB and 0.2 ms on CNUH, where NeuroKit2 measures most beats. A comparison restricted to commonly measured beats therefore favors every method evaluated on it, including QTANet, and is reported here alongside the unconditional figures.

3.6. Attention and External T-Wave Delineation

Table 9 summarizes the recording-level frame comparison. Manual and automatic QTDB reference analyses are reported separately because they cover different subsets and use different seed sets.

Table 9. Recording-level frame-metric comparison, QTANet versus Control (–Attention). *p*-values are unadjusted.

Reference	Dataset	Metric	<i>n</i> Rec	QTANet	Control	Δ	Wilcoxon <i>p</i>	QTANet Better
Expert .q1c	QTDB	T-wave F1	41	0.740 ± 0.159	0.678 ± 0.229	+0.061	2.0×10^{-5}	73.2%
Expert .q1c	QTDB	T-wave AUC	41	0.932 ± 0.076	0.907 ± 0.115	+0.024	7.3×10^{-6}	80.5%
Expert .q1c	QTDB	Accuracy	41	0.838 ± 0.094	0.825 ± 0.101	+0.013	0.013	73.2%
Expert .q1c	QTDB	QRS F1	41	0.807 ± 0.118	0.813 ± 0.128	−0.006	0.062	29.3%
Automatic	QTDB	T-wave F1	105	0.7713	0.6900	+0.0813	2.9×10^{-16}	90.5%
Automatic	QTDB	T-wave AUC	105	0.9385	0.9037	+0.0348	2.0×10^{-16}	91.4%
Automatic	QTDB	Macro F1	105	0.8270	0.7898	+0.0372	4.6×10^{-15}	89.5%

Table 9. Cont.

Reference	Dataset	Metric	<i>n</i> Rec	QTANet	Control	Δ	Wilcoxon <i>p</i>	QTANet Better
Expert	LUDB (OOF)	T-wave F1	160	0.8896	0.8649	+0.0247	4.8×10^{-14}	78.1%
Expert	LUDB (OOF)	Macro F1	160	0.9154	0.8986	+0.0167	5.8×10^{-18}	85.0%
Expert	CNUH	T-wave F1	33	0.7988	0.7872	+0.0117	0.012	75.8%

Per-recording metrics are averaged over seeds 123, 7, 41 and 42 except the expert (.q1c) block, which uses seeds 123, 41 and 42. Differences use unrounded values. The expert QTDB block contains 41 fully bracketed recordings, 177 windows and 102,495 T-wave frames. Record-averaged AUC differs from frame-pooled AUC because recording weights differ. The different reference/seed subsets limit direct comparisons between blocks.

Against expert annotations, attention improves T-wave F1 in 30 of 41 recordings (0.740 versus 0.678; $p = 2.0 \times 10^{-5}$) and T-wave AUC in 33 of 41 (0.932 versus 0.907; $p = 7.3 \times 10^{-6}$). The automatic-reference analysis also favors QTANet. This consistency supports the direction of the T-wave result, although the analyses use different recordings, annotations and seed sets and are not interchangeable measurements.

The detection and frame-level analyses provide complementary observations. Against the automatic reference, QTDB PPV improves for both QRS and T offsets, while T-offset sensitivity also increases. In the expert frame-level subset, T-wave F1 improves but QRS F1 does not show a corresponding increase (difference -0.006 ; $p = 0.062$). This asymmetry may partly reflect the length of the spurious segments: a spurious QRS segment counts as a full false positive in a metric that counts waves but contributes proportionally fewer samples in a metric that weights samples, because spurious QRS segments are short (73 ms on average, against 112 ms for true-positive segments; Section 3.10).

False-positive waveform predictions can generate incorrect intervals, whereas missing landmarks can lead to missing measurements. Both are relevant to monitoring applications. Joint reporting of detection, interval error and coverage is therefore needed; these results do not establish that one failure type is uniformly less consequential clinically.

The 41 fully bracketed records have lower observed automatic-reference T-wave F1 than the remaining 64 records (0.741 versus 0.796) and lower accuracy (0.830 versus 0.867). The reported Mann–Whitney *p*-values range from 0.08 to 0.25. Thus, there is no observed performance advantage for the selected subset on those metrics, but nonsignificance does not demonstrate equivalence or eliminate selection bias.

The expert-reference T-wave F1 gain is smaller internally (+0.0247) and on CNUH (+0.0117) than on QTDB (+0.061). The clinical direction varies across seeds. These results support a dataset-dependent association between attention and T-wave delineation, without establishing uniform interval-level superiority or a specific cause for the between-cohort differences.

An additional analysis of received attention weights reports a mean ratio of 1.56 (median 1.61) for attention received within ± 40 ms of annotated boundaries relative to the 100–300 ms bands in 19 held-out recordings (Figure 7). The ratio exceeds one in all 19 recordings (Wilcoxon $p = 3.8 \times 10^{-6}$). This describes alignment of attention with annotated landmarks; it does not validate attention weights as a causal explanation or establish the clinical usefulness of inspecting them.

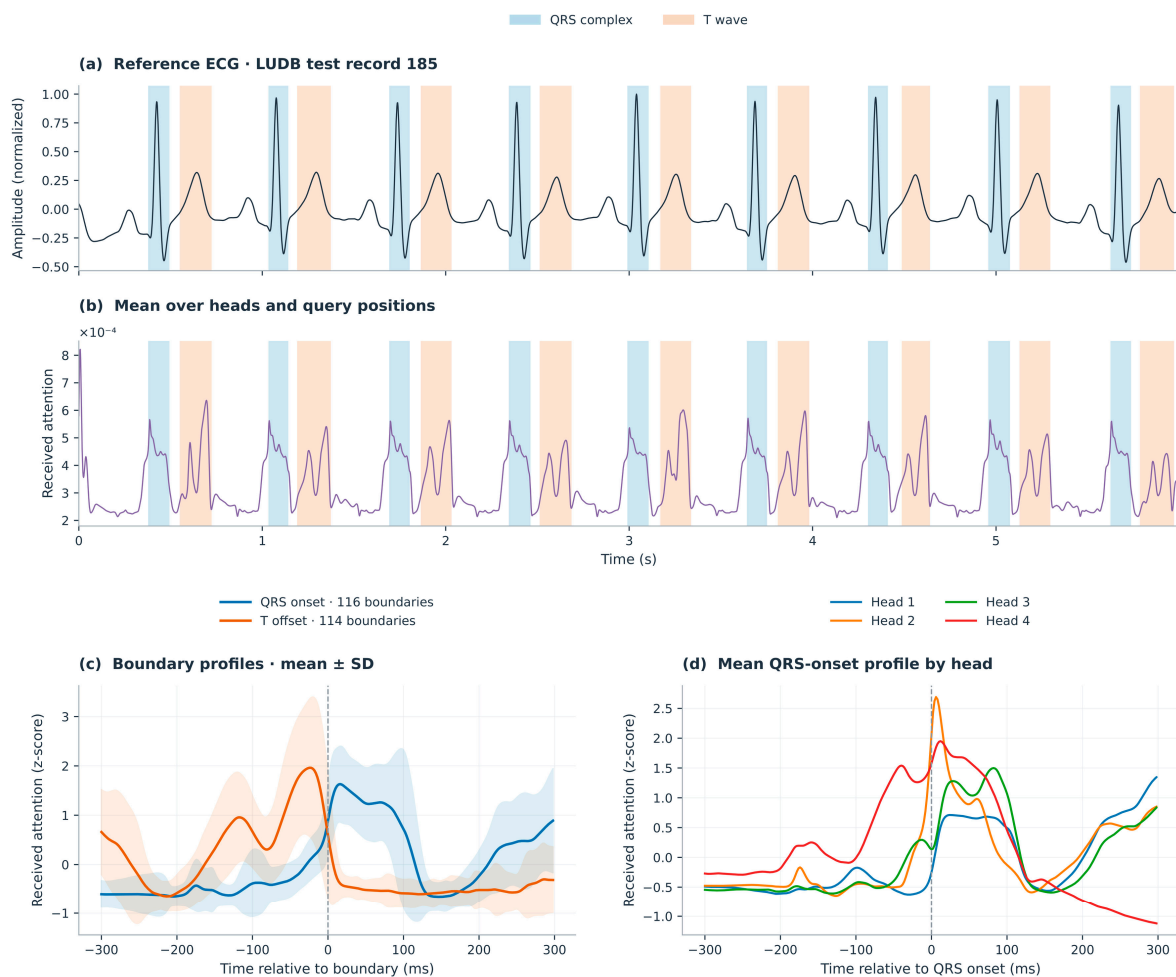


Figure 7. Received self-attention. (a) LUDB record 185 with reference segments; (b) attention averaged over heads and query positions; (c) boundary-aligned z-score profiles, mean $\pm$ SD, for 116 QRS onsets and 114 T offsets; (d) mean QRS-onset profiles by head. These descriptive patterns do not establish a causal explanation of predictions.

3.7. Ablation Study

Table 10 and Figure 8 describe the component ablation. Table 10 reports QTDB results as means $\pm$ SD across four training seeds, each the mean of five per-fold models; Figure 8 shows the seed-123 runs for LUDB OOF, QTDB and CNUH. Removing attention produces the largest loss in T-wave F1 (from 0.7699 ± 0.0139 to 0.7150 ± 0.0194), about four times the seed-to-seed SD of the full model. The full model has the highest accuracy and T-wave F1 and the lowest QT MAE of all variants; removing the FFN gives a marginally higher macro F1 (83.14% versus 82.97%), well within one seed-to-seed SD. No single component dominates the QT error. The largest single-component increase in QT MAE is that of the second Bi-LSTM (+6.85 ms). Removing the DWT branch, which the seed-123 run identified as the most costly removal (+17.18 ms), costs about 4.3 ms across four seeds; extended to eight seeds, the cost falls to 1.78 ms (37.22 ± 3.06 versus 39.00 ± 4.45 ms), within the seed-to-seed SD. The seed-123 figure arose because that seed was the best of the four for the full model and the worst for the $-$ DWT variant, so a reliable individual contribution of the DWT branch is not established at this sample size. This is consistent with the wavelet sensitivity analysis (Section 2.4), in which accuracy varies by 1.74 percentage points across nine configurations. Because the eight-seed extension covered only the full model and the $-$ DWT variant, the QT-MAE margins of the other variants rest on four seeds; for attention, the stronger evidence is the T-wave F1 difference. Internally, the comparison of QTANet

with –Attention depends on the evaluation set: –Attention has lower beat-pooled QT MAE on the 19-recording fixed test set (Table 6), whereas on the 160-recording out-of-fold set averaged over four seeds, QTANet is better in macro F1 (91.69% versus 90.16%) and QT MAE (18.42 versus 20.13 ms; Section 3.3). Control and – Attention denote the same configuration.

Table 10. Component ablation on QTDB (automatic reference; M2, 50–700 ms). Mean $\pm$ SD across four training seeds, each the mean of five per-fold models.

Variant	Accuracy (%)	Macro F1 (%)	T-Wave F1	QT MAE (ms)
QTANet (full)	85.29 $\pm$ 1.04	82.97 $\pm$ 1.34	0.7699 $\pm$ 0.0139	35.29 $\pm$ 3.02
–FFN	85.19 $\pm$ 0.48	83.14 $\pm$ 0.56	0.7603 $\pm$ 0.0077	39.29 $\pm$ 1.54
–Attention (Control)	82.87 $\pm$ 1.19	79.37 $\pm$ 1.51	0.7150 $\pm$ 0.0194	39.46 $\pm$ 1.43
–2nd Bi-LSTM	84.30 $\pm$ 0.46	82.24 $\pm$ 0.22	0.7524 $\pm$ 0.0120	42.14 $\pm$ 1.29
Single-head attention	84.12 $\pm$ 1.17	81.46 $\pm$ 1.58	0.7511 $\pm$ 0.0220	39.68 $\pm$ 3.29
–DWT	84.42 $\pm$ 1.65	81.92 $\pm$ 1.91	0.7500 $\pm$ 0.0278	39.58 $\pm$ 6.52

QT error is derived through the common post-processing rule, and dispersion is across seeds, not folds. The eight-seed extension covers the full model and the –DWT variant only (QT MAE 37.22 $\pm$ 3.06 versus 39.00 $\pm$ 4.45 ms). Single-seed (seed 123) results for LUDB OOF, QTDB and CNUH are shown in Figure 8. The table describes the evaluated runs and does not establish a general ranking of architectural importance.

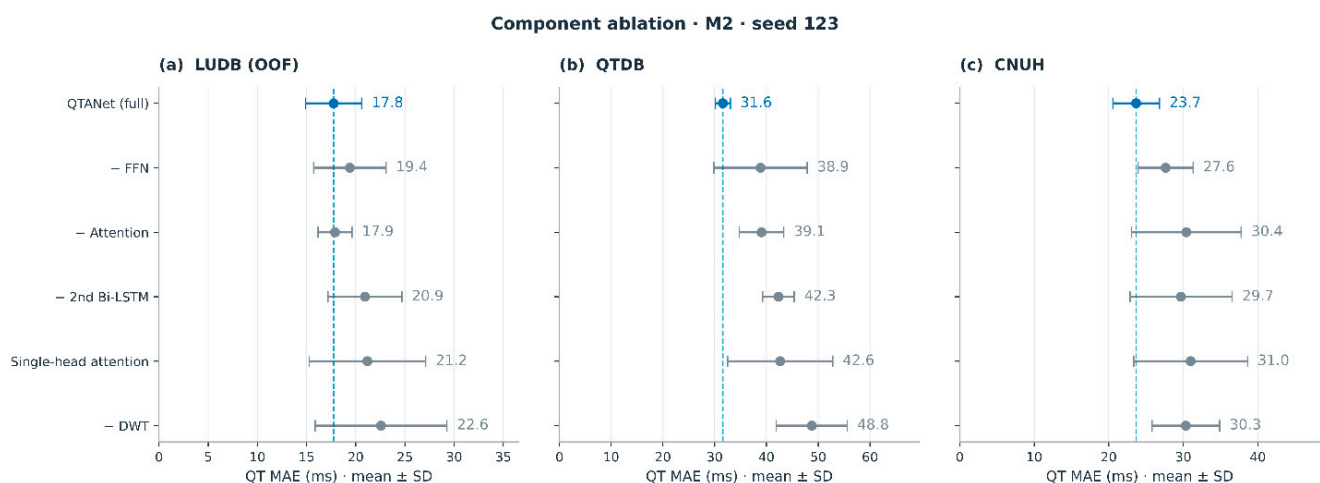


Figure 8. Component ablation under M2 (50–700 ms), seed 123. Points and horizontal error bars show mean QT MAE and SD across folds. (a) LUDB OOF; (b) QTDB; (c) CNUH. Dashed vertical lines mark full-model means. The full model is shown in blue and the ablated variants in gray. These single-seed comparisons are descriptive; Table 10 reports the four-seed QTDB summary, in which the large QTDB cost of removing the DWT branch in this run is not reproduced.

To test whether the placement of the attention stage matters, three variants were trained that differ from QTANet only in where the attention sub-layer sits, with its internal configuration (four heads, the feed-forward sub-layer, the normalization placement and dropout), the loss, the focal schedule, the fold assignment, early stopping and post-processing held fixed: before the first Bi-LSTM, applied to the 384-dimensional fused features either after a linear projection to 128 dimensions (9.84 M parameters) or directly at 384 dimensions with $d_k = 96$ (10.79 M), and after the second Bi-LSTM, immediately before the classifier (9.74 M). Each variant was trained with the same four seeds and five folds. The retained placement has the highest accuracy, macro F1 and T-wave F1 and the lowest QT MAE in all three cohorts. On QTDB, its QT MAE is 35.29 $\pm$ 3.02 ms, against 43.57 $\pm$ 2.74, 44.72 $\pm$ 2.80 and 42.20 $\pm$ 0.85 ms for the three variants (paired t-tests across the four seeds, $p = 0.015$, 0.020 and 0.022). On LUDB OOF, the differences are smaller (18.42 $\pm$ 1.13 ms against 21.78 $\pm$ 1.82, 22.32 $\pm$ 0.66 and 20.06 $\pm$ 0.90 ms), and the placement after the second

Bi-LSTM does not differ significantly ($p \geq 0.098$); on CNUH, with 33 recordings, most differences are not significant. The placement after the second Bi-LSTM has exactly the same parameter count as QTANet, so its deficit cannot be attributed to capacity, and the two placements before the first Bi-LSTM perform alike, although they add 0.10 M and 1.05 M parameters. In the error decomposition of Section 3.10, moving the attention stage increases QRS segments predicted over reference T waves 1.7- to 4.7-fold relative to QTANet, against 1.5- to 1.9-fold for segments overlapping no annotated wave (per window, means of four seeds). This sensitivity analysis compares three alternative placements only; it supports the retained placement among them but does not establish that it is optimal.

Figure 9 illustrates residual boundary errors despite generally accurate waveform delineation in a selected LUDB segment. QTANet achieved higher QRS and T-wave F1 scores and a slightly lower QT mean absolute error than the control model in this segment. Nevertheless, in the highlighted beat, QTANet predicted the QRS onset 10 ms early and the T-wave offset 20 ms late, overestimating the QT interval by 30 ms. This qualitative example does not establish the frequency of these errors or overall model superiority.

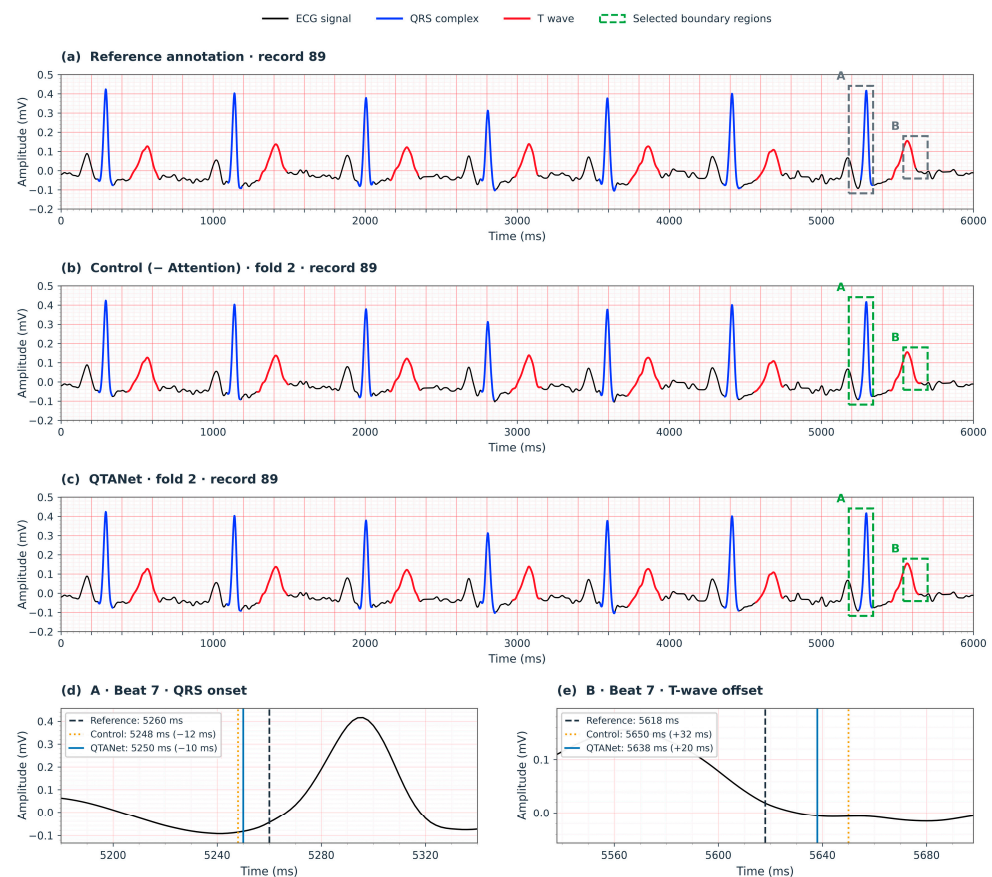


Figure 9. Residual delineation errors in a LUDB segment (recording 89, lead II, 6 s). (a) Expert annotation; (b) model without self-attention; (c) QTANet. Dashed boxes A and B (gray in (a), green in (b,c)) mark the regions enlarged in (d,e), showing QRS onset and T-wave offset errors, respectively. Recording 89 was selected as an illustrative example and is not among the recordings with the largest QTANet QT error (11th of 19 on the fixed test set for seed 123; QT MAE 10.7 ms); it shows that small boundary displacements can alter the derived interval even when frame-level delineation is visibly accurate (Section 4.2).

3.8. Loss-Schedule Sensitivity

Table 11 compares six loss configurations on the LUDB fixed test set as means $\pm$ SD across four training seeds, each the mean of five per-fold models; Figure 10 shows the seed-

123 runs of five of these configurations, excluding the budget-matched control described below. Because γ increases over a nominal 200-epoch budget, whereas early stopping halts training when the validation loss stops improving, the configurations reach different effective values of γ and train for different lengths: for seed 123, the 0-to-2 schedule reaches $\gamma = 0.65$ at the selected checkpoint and stops after 75.6 epochs on average, whereas the 0-to-5 schedule reaches 3.50 and stops after 150.2 epochs. The comparison therefore concerns the training procedures rather than the schedules alone. A budget-matched control ramps γ from 0 to 2 over 80 epochs, so that the nominal endpoint falls within the observed training length. The superiority of the 0-to-5 schedule is reproducible across seeds: it has the highest accuracy (94.48%), macro F1 (92.53%) and T-wave F1 (0.9006) and the lowest QT MAE (15.39 ms). At the same nominal endpoint of $\gamma = 2$, the budget-matched ramp outperforms the fixed schedule by 1.4 macro-F1 points (91.84% versus 90.43%), consistent with the curriculum rationale in Section 2.7, while their QT errors are similar (18.29 versus 18.13 ms). Matching the budget improves the QT MAE of the 0-to-2 ramp from 21.24 to 18.29 ms, which partly explains its lower performance relative to the 0-to-5 schedule, although it does not reach that performance. The main 0-to-2 configuration is thus one of the evaluated configurations, not the empirically best one. It was fixed before the experiments and is retained, because adopting the better-performing schedule on the basis of fixed-test-set results would constitute test-set selection. These sensitivity results are not accompanied by a paired significance test for schedule superiority.

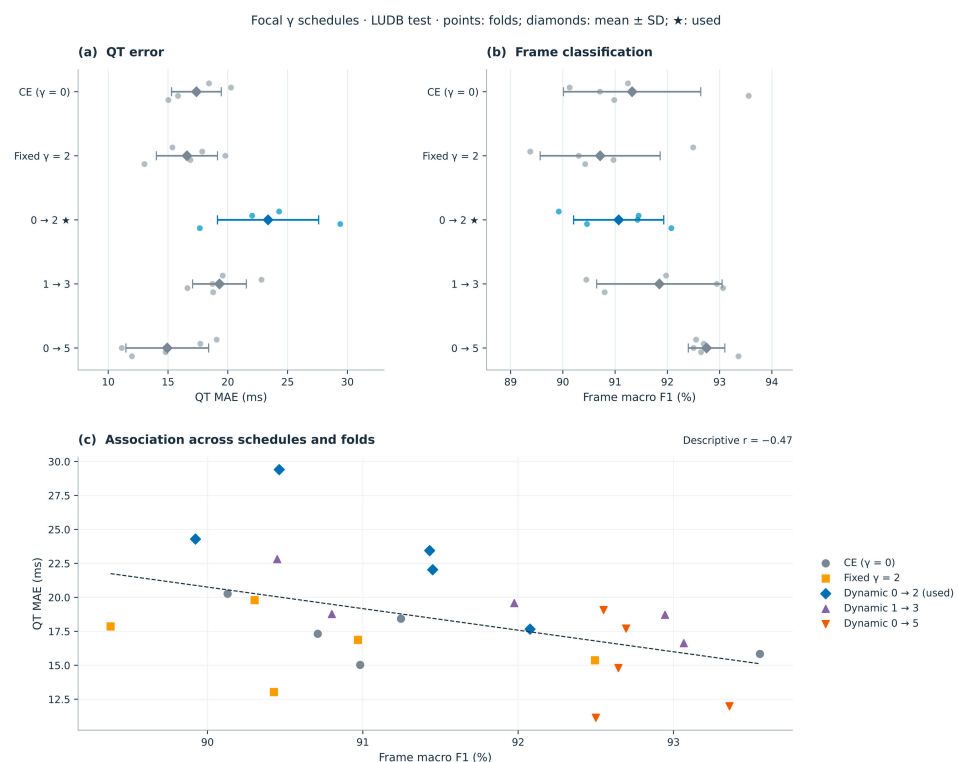


Figure 10. Focal-loss schedule sensitivity on the LUBD fixed test set (seed 123; five per-fold models). (a) QT MAE; (b) frame macro F1. Small points represent folds, diamonds show means and horizontal bars show SD. The star identifies the 0-to-2 schedule used in the main experiments. In (a,b), this schedule is shown in blue and the other schedules in gray. (c) QT MAE versus frame macro F1 across schedules and folds; the dashed line illustrates the descriptive association ($r = -0.47$). Shared folds and evaluation recordings preclude treating the plotted observations as independent replicates; no independent-sample p -value is reported. The four-seed summary, including the budget-matched control, is given in Table 11.

Table 11. Focusing-parameter schedule sensitivity (LUDB fixed test set; mean $\pm$ SD across four seeds, each the mean of five folds).

Configuration	Accuracy (%)	Macro F1 (%)	T-Wave F1	QT MAE (ms)
Cross-entropy ($\gamma = 0$)	93.22 $\pm$ 0.31	90.95 $\pm$ 0.49	0.8750 $\pm$ 0.0070	18.80 $\pm$ 1.04
Fixed $\gamma = 2$	92.82 $\pm$ 0.22	90.43 $\pm$ 0.38	0.8673 $\pm$ 0.0038	18.13 $\pm$ 1.82
Dynamic 0 $\rightarrow$ 2 (used in this paper)	93.60 $\pm$ 0.41	91.59 $\pm$ 0.53	0.8795 $\pm$ 0.0082	21.24 $\pm$ 2.35
Dynamic 1 $\rightarrow$ 3	93.44 $\pm$ 0.22	91.40 $\pm$ 0.35	0.8785 $\pm$ 0.0043	20.87 $\pm$ 2.24
Dynamic 0 $\rightarrow$ 5	94.48 $\pm$ 0.22	92.53 $\pm$ 0.35	0.9006 $\pm$ 0.0026	15.39 $\pm$ 0.53
Dynamic 0 $\rightarrow$ 2, γ budget of 80 epochs	93.89 $\pm$ 0.22	91.84 $\pm$ 0.34	0.8884 $\pm$ 0.0033	18.29 $\pm$ 2.59

Dispersion is across seeds, not folds. The budget-matched control ramps γ from 0 to 2 over 80 epochs, so that the nominal endpoint falls within the observed training length. For seed 123, the γ reached at the selected checkpoint and the mean stopping epoch are 0.00 and 50.6 for cross-entropy, 2.00 and 43.0 for fixed $\gamma = 2$, 0.65 and 75.6 for 0 $\rightarrow$ 2, 1.61 and 71.6 for 1 $\rightarrow$ 3, and 3.50 and 150.2 for 0 $\rightarrow$ 5. Because these values refer to the fixed 19-recording test set, the QT MAE of the 0 $\rightarrow$ 2 schedule is not directly comparable with the out-of-fold value of the same configuration in the class-weight comparison (18.42 ms; Section 2.7).

Figure 10c shows a descriptive association between frame macro F1 and QT MAE across schedules and folds ($r = -0.47$). Because configurations share folds and evaluation recordings, this coefficient is not accompanied by an independent-sample p -value and does not establish a general predictive relationship between the metrics. Both frame-level and interval-level outcomes are needed to evaluate the measurement pipeline.

3.9. Computational Complexity

Table 12 reports model size, FLOPs, memory, training time and inference latency, measured on the hardware used throughout. QTANet has 9.74 M parameters, 62.91 GFLOPs per window and a batch-1 latency of 10.33 ms per 6-s window on the GPU, about 580 times faster than real time, and 200.1 ms on the CPU; the control has 9.60 M parameters and latencies of 8.90 and 162.2 ms. Of the QTANet FLOPs, 53.09 G (84.4%) are spent in the residual encoder, which operates at the full input length of 3000 samples; the recurrent and attention stack accounts for 7.95 G (12.6%), and self-attention for 5.00 G (8.0%). The computational cost of the design lies mainly in the absence of downsampling; attention accounts for a smaller share. Attention nevertheless roughly doubles the batch-1 inference peak memory, from 119.1 to 234.6 MB; about 34 MB of this increase is a head-averaged attention map that is materialized only because attention weights are returned by default, and the remainder reflects the 3000×3000 attention maps, whose quadratic scaling with sequence length constrains longer windows. The modern U-Net variant is substantially cheaper, particularly on the CPU (10.8 ms per window), which is the relevant figure for bedside or ambulatory deployment. The timings indicate window-processing throughput on this hardware; they exclude signal acquisition and do not establish end-to-end streaming latency.

FLOPs are counted analytically at two FLOPs per multiply–accumulate (batch 1, input length 3000) over all convolutional, linear, recurrent and attention layers, and cross-checked against PyTorch FlopCounterMode with the fused attention path disabled; automatic profilers report zero for the fused attention and LSTM kernels used at inference. Inference peak memory is the batch-1 peak allocation without gradients, including the model weights. Training time per fold is the mean $\pm$ SD, with the mean number of epochs in parentheses. Latency is the mean $\pm$ SD over 50 GPU runs and 10 CPU runs on an otherwise idle device and excludes the CPU-side wavelet transform (0.40 ms per window), window acquisition and the complete measurement pipeline. These timings should not be extrapolated to other hardware without measurement.

Table 12. Computational cost (input: one 6-s window of 1×3000 samples; batch 1; NVIDIA RTX 4090 GPU and 8-thread CPU).

Model	Parameters (M)	FLOPs (G)	Checkpoint (MB)	Inference Peak Memory (MB)	Training per Fold (s)	GPU Latency (ms)	CPU Latency (ms)
QTANet	9.74	62.91	37.3	234.6	142 ± 33 (81)	10.33 ± 0.23	200.1 ± 0.4
–Attention (Control)	9.60	57.52	36.8	119.1	108 ± 29 (74)	8.90 ± 0.19	162.2 ± 0.4
–DWT	9.36	60.65	35.8	227.3	103 ± 80 (57)	9.54 ± 0.27	192.3 ± 0.9
ResNet + LSTM + DWT	9.50	56.93	36.4	118.8	102 ± 34 (74)	7.01 ± 0.15	136.9 ± 3.2
U-Net, modern variant	4.10	3.59	15.7	31.2	40 ± 6 (41)	2.40 ± 0.01	10.8 ± 1.7
U-Net, published	0.167	0.11	0.69	9.4	137 ± 49 (113)	1.68 ± 0.01	3.05 ± 0.01

3.10. Decomposition of False-Positive Segments

To locate the spurious waveform predictions behind the PPV differences in Section 3.1, each predicted QRS and T-wave segment was assigned to one of four categories according to the reference wave it overlaps: a reference wave of the same type (a true positive), a reference wave of a different type, a reference P wave, or no annotated wave. P-wave positions were recovered from the QTDB annotations. Counts are summed over the five per-fold models and reported for seed 123 in Table 13; the decomposition was repeated for seeds 7, 41 and 42.

Table 13. False-positive segments on QTDB by the reference wave they overlap (automatic reference; seed 123; counts summed over the five per-fold models).

False-Positive Type	Control (–Attention)	QTANet	Reduction	Modern U-Net (4.10 M)	Published U-Net (0.167 M)
QRS predicted over a reference T wave	52,712	4913	90.7%	14,432	52,798
T wave predicted over a reference P wave	37,150	8688	76.6%	12,098	23,229
Segment overlapping no annotated wave	135,351	60,784	55.1%	69,807	91,773

Reductions are relative to Control (–Attention). P-wave positions were recovered from the QTDB annotations. The U-Net comparators are described in Sections 2.9 and 3.11; their QT detection rates on QTDB are 94.5% (modern) and 68.3% (published), against 92.8% for QTANet. Across seeds 123, 7, 41 and 42, the reductions range from 55.0% to 90.7% for QRS predicted over a reference T wave and from 21.3% to 55.1% for segments overlapping no annotated wave; for T waves predicted over a reference P wave, the change ranges from a 76.6% reduction to a 92.0% increase.

On QTDB, attention reduces false-positive segments by 78.7% for QRS and by 57.8% for T waves. For seed 123, Table 13 shows that the proportional reduction is largest for confusions of wave identity: QRS segments predicted over reference T waves decrease from 52,712 to 4913 (90.7%), and T-wave segments predicted over reference P waves from 37,150 to 8688 (76.6%), compared with a reduction of 55.1% for segments overlapping no annotated wave. Across the four training seeds, the reduction in QRS segments predicted over reference T waves is consistent (55.0–90.7%) and in every seed exceeds the reduction for segments overlapping no annotated wave (21.3–55.1%). The change in T-wave segments predicted over reference P waves is not stable across seeds, ranging from a 76.6% reduction to a 92.0% increase, and no individual contribution is claimed for that category. The same direction holds on the out-of-fold LUDB recordings, at a much smaller magnitude (QRS false positives per window, 0.4875 to 0.2875), and on CNUH (0.2485 to 0.1030). Spurious QRS segments are short, 73 ms on average against 112 ms for true-positive QRS segments, which may partly explain why the effect appears in QRS PPV but not in QRS frame-level F1 (Section 3.6).

This analysis characterizes the error patterns associated with the attention module rather than establishing a causal mechanism. Without attention, a large share of the spurious predictions confuse wave identity, with P waves read as T waves and T waves read as QRS complexes; the convolutional encoder is local, and the recurrent stages integrate context sequentially rather than by directly weighting the other beats in the window. One possible interpretation is that self-attention allows each time step to draw on the whole window, so that a deflection can be classified in the context of the surrounding rhythm; this interpretation is consistent with the expert-reference T-wave results on the 41 fully bracketed QTDB recordings (Section 3.6). For seed 123, QTANet also produces fewer false-positive segments on QTDB than either U-Net comparator (Section 3.11) in every category of Table 13. The suppression of spurious waves is not unique to self-attention, however. The modern U-Net, which has skip connections and no attention, produces fewer spurious QRS segments per window than QTANet on LUDB (0.250) and CNUH (0.067), and fewer spurious T-wave segments on LUDB (0.206 versus 0.244). On QTDB, the advantage of QTANet depends on the training seed: its rate of spurious QRS segments varies about threefold across the four seeds (0.29–0.89 per window), whereas that of the modern U-Net varies little (0.38–0.44), so that the ordering reverses when the rates are averaged over seeds. The analysis therefore does not show that attention is the only route to fewer spurious waves.

Whether the reduction in wave-identity confusion is related to the measurement task was examined across the 104 QTDB recordings, with all quantities averaged over the four training seeds before pairing. The per-window reduction, relative to Control, in segments overlapping a reference wave of a different type or a reference P wave correlates with the recording-level reduction in QT MAE (Spearman $\rho = 0.37$, $p = 1.2 \times 10^{-4}$) and, more strongly, with the gain in the proportion of reference beats measured ($\rho = 0.62$, $p = 2.9 \times 10^{-12}$). The reduction in segments overlapping no annotated wave is not associated with that gain ($\rho = 0.06$, $p = 0.52$), and the association of identity confusion with coverage persists after partialling it out ($\rho = 0.64$, $p = 2.8 \times 10^{-13}$). In the third of recordings with the largest reduction in identity confusion, QT MAE improves by 13.5 ms, against 2.9 ms in the remainder (Mann–Whitney $p = 7.2 \times 10^{-5}$), and coverage by 11.8 against 0.4 percentage points ($p = 5.5 \times 10^{-11}$). Both false-positive types are, however, similarly associated with QT MAE itself ($\rho = 0.37$ and 0.36), so the specific association is with measurement coverage rather than with interval accuracy. These associations across recordings do not establish a causal pathway.

3.11. Comparison with Published Deep-Learning Delineators

Table 14 compares QTANet with the reimplemented one-dimensional U-Net of Moskalenko et al. [17] (Section 2.9), trained under an identical protocol. QTANet is superior on both external cohorts in frame-level metrics: macro F1 is higher in 101 of 105 QTDB recordings ($\Delta = +0.0742$; $p = 8.6 \times 10^{-18}$) and in all 33 CNUH recordings ($\Delta = +0.0615$; $p = 2.3 \times 10^{-10}$), and T-wave AUC is higher on both cohorts. QT MAE on QTDB is also lower ($\Delta = -5.40$ ms; $p = 0.025$), whereas the QTDB T-wave F1 difference is not significant ($p = 0.094$). On the internal out-of-fold set, in contrast, the U-Net is better in T-wave F1 ($\Delta = -0.0159$; $p = 1.2 \times 10^{-6}$) and QT MAE ($\Delta = +1.35$ ms; $p = 0.0037$). This asymmetry is consistent with a benefit of the proposed design that is specific to distribution shift rather than general, although the published U-Net is also much smaller than QTANet (0.167 M versus 9.74 M parameters). On LUDB, its T-wave F1 is the highest among the compared models (0.905), whereas its macro F1 is the lowest (0.862), indicating weaker QRS delineation.

Table 14. Per-recording paired comparison of QTANet with the published one-dimensional U-Net [17] trained under an identical protocol. Metrics were averaged over four seeds before testing; *p*-values are unadjusted.

Dataset	Metric	Δ (QTANet – U-Net)	Wilcoxon <i>p</i>	Recordings Favoring QTANet
QTDB	Macro F1	+0.0742	8.6×10^{-18}	101/105
QTDB	T-wave AUC	+0.0475	5.2×10^{-14}	96/105
QTDB	QT MAE (ms)	−5.40	0.025	63/104
QTDB	T-wave F1	+0.0270	0.094	52/105
CNUH	Macro F1	+0.0615	2.3×10^{-10}	33/33
CNUH	T-wave AUC	+0.0193	2.0×10^{-6}	30/33
LUDB (OOF)	T-wave F1	−0.0159	1.2×10^{-6}	52/160
LUDB (OOF)	QT MAE (ms)	+1.35	0.0037	53/160

For QT MAE, a negative Δ favors QTANet. QTDB uses the automatic reference and compares QT MAE on 104 recordings (Section 3.2); LUDB and CNUH use expert annotations.

The larger modern U-Net variant (4.10 M parameters), trained under the same protocol, outperforms QTANet in frame-level metrics, with macro F1 higher by 0.020 on QTDB ($p = 7.7 \times 10^{-9}$), whereas its QT error does not differ significantly from that of QTANet on LUDB ($p = 0.28$) or CNUH ($p = 0.43$). On QTDB, its QT MAE is lower than that of QTANet by 3.06 ms ($p = 5.6 \times 10^{-5}$; QTANet has the lower error in 30 of 104 recordings), although this difference depends on the seed: for seed 123 alone, it reverses (QTANet lower by 4.05 ms; $p = 0.83$; QTANet has the lower error in 50 of 104 recordings). The comparison therefore supports the superiority of QTANet over the published delineator under distribution shift, not architectural optimality. The modern U-Net also attains a tighter QRS-onset error distribution on QTDB (SD 15.83 versus 17.51 ms for QTANet, both averaged over four seeds; 15.70 versus 17.85 ms for seed 123, the basis of Table 2), although its systematic QRS-onset bias is larger (+6.48 versus +5.78 ms, four-seed means), so the boundary precision of QTANet is not attributed to the full-resolution design of its encoder.

4. Discussion

4.1. Principal Findings and Comparison with Existing Approaches

QTANet estimates QT from predicted waveform boundaries, and its comparative performance depended on the dataset, reference, comparator, and evaluation subset. In the expert-annotated QTDB interval subset, it had lower QT MAE and higher measurement coverage than the comparators listed in Table 6. Expert-reference T-wave frame metrics also favored QTANet over the attention-free Control in the subset permitting complete waveform labeling. These findings support reporting interval error and coverage alongside frame-level performance; they do not establish uniform gains across evaluation settings or training runs.

Under the matched training and evaluation protocol, QTANet achieved higher macro F1 and T-wave AUC in both external cohorts and lower QTDB QT error than the reimplemented one-dimensional U-Net of Moskalenko et al. [17]. However, the published U-Net had higher T-wave F1 and lower QT error in the LUDB out-of-fold (OOF) evaluation. The modern U-Net had 4.10 million parameters, compared with 0.167 million for the published U-Net and 9.74 million for QTANet. Despite having no attention module, it achieved stronger frame-level performance and lower four-seed mean QT error on QTDB than QTANet. The contribution is therefore not general architectural superiority, but a characterization of waveform-identity errors and their relationship to interval accuracy and measurement coverage.

The QTDB window-level regression comparison used QT references derived from automatically generated ecgpuwave annotations for the first ECG channel (.pu0 files),

rather than the expert annotations used in the separate interval analysis. On common valid windows, QTANet had lower QT MAE than the implemented scalar regression baseline under the evaluated training budget, with recording-level bootstrap confidence intervals favoring QTANet on all three datasets. The comparison averaged QTANet absolute errors across four seeds before pairing them with one regression run; it was neither seed-matched nor a prediction ensemble. Dense supervision is a possible contributor to the observed difference, but its effect was not separated from architecture or optimization. Published QTNet and QTcNet studies used substantially larger ECG datasets [12,13]; differences in endpoints, references, populations, and training scale preclude direct ranking with the present results.

4.2. Attention, Delineation, and QT Accuracy

Frame discrimination and QT accuracy are related but non-equivalent. Frame-level metrics assess waveform classification, whereas QT depends on paired boundary locations. For a matched beat, the signed QT error equals the T-offset error minus the QRS-onset error; better classification within a waveform therefore need not improve interval accuracy. On the fixed 19-recording LUDB test set, Control and ResNet + LSTM + DWT had lower beat-pooled QT MAE than QTANet (17.76 and 16.53 versus 23.37 ms; Table 6). The QTANet–Control gap persisted after symmetric trimming and was predominantly associated with T-offset error. This result differs from the small advantage for QTANet over Control in the LUDB OOF analysis. The evaluation sets and aggregation procedures differ, so these results are not interchangeable estimates.

In the automatic-reference QTDB analysis, attention was associated with higher positive predictive value for QRS onset and T offset, a small reduction in QRS-onset sensitivity, and increased T-offset sensitivity. Across four training seeds, QTANet produced 55.0–90.7% fewer QRS segments overlapping reference T waves than Control, a larger reduction in every seed than for segments overlapping no annotated wave (21.3–55.1%); the change in T-wave segments overlapping reference P waves was not consistent across seeds. Recordings with larger reductions in wave-identity confusion showed larger gains in measurement coverage, and moving the attention stage from its position between the recurrent stages increased these confusions more than other false positives (Sections 3.7 and 3.10). These patterns are consistent with fewer waveform-identity confusions, but neither the error decomposition nor attention-map alignment establishes a causal mechanism. The attention-free modern U-Net also reduced these false-positive categories relative to Control, indicating that such reductions are not specific evidence for an attention-based mechanism.

Multi-seed ablation provided stronger evidence for an attention-related T-wave F1 improvement than for a uniform QT benefit. All four QTDB interval point estimates favored QTANet, but the seed-42 difference was small and nonsignificant. The main ablation covered four seeds; only the full model and the no-DWT variant were extended to eight. The reduced full-versus-no-DWT difference in that extension leaves the magnitude and stability of the DWT contribution uncertain, rather than establishing either a consistent benefit or no effect. Component-level conclusions should therefore distinguish frame-level improvements from interval-level effects and their variability across training runs.

4.3. Measurement Coverage and Generalizability

Reported QT MAE was conditional on a method returning a valid, matched measurement. QTANet generally had higher measurement coverage than NeuroKit2, making joint interpretation of coverage and error essential. The common-window regression analysis does not by itself quantify success on all reference-eligible windows; missing or invalid references and model measurement failures require separate accounting. The 50–700 ms range

is an evaluation filter, not a physiological limit, and can exclude genuinely prolonged QT intervals. Under automatic QTDB references, widening the range to 50–1000 ms increased coverage from 92.8% to 96.2% and MAE from 31.58 to 39.50 ms, while signed bias changed from -7.26 to -15.19 ms. Because the admitted reference intervals include both genuine QT prolongation and automatic-annotation errors, this shift cannot be attributed solely to model underestimation.

The interval evaluation is reference-anchored. Reference QRS midpoints select predicted segments for scoring, while the QRS onset and T offset are model-derived. Reference annotations are not network inputs, but the matching procedure determines which predicted beats enter the reported interval analyses. The resulting accuracy and coverage are therefore not equivalent to reference-free end-to-end performance. Lead configuration also matters: LUDB and CNUH used lead II, whereas QTDB used its first channel, and the QTDB expert annotators had access to both leads. Differences in T-wave amplitude and terminal slope across leads may affect boundary measurement, and no lead-stratified comparison under a common protocol was performed.

Generalizability is bounded by the realized cohorts. A selection-list-to-file mapping error left 18 LUDB recordings outside the four intended sinus-rhythm classes, including 16 with atrial fibrillation or flutter and two with irregular sinus rhythm. The reported results describe this realized cohort and do not replace rhythm-stratified validation. CNUH comprised 33 recordings from 33 patients and 209 reference beats, acquired at one site with one device and annotated by one cardiologist. Its T-wave F1 gain was small and seed-dependent; this cohort does not establish sensitivity or specificity for QT prolongation or multicenter generalizability. The expert-reference QTDB analyses also used restricted, endpoint-specific subsets. Broader validation is needed across rhythms, morphologies, prolonged-QT populations, and acquisition settings.

4.4. Limitations and Future Directions

The development pool contained 160 recordings, with 128 used for optimization in each standard fold. The separate fixed test set contained only 19 recordings and was sensitive to outliers. LUDB OOF estimates were non-nested because the held-out fold also informed early stopping, introducing model-selection exposure. The nested sensitivity analysis addressed frame-level metrics only and did not validate the interval-level results under a nested protocol. Correlated frames within a recording should not be treated as additional independent observations. The two-seed learning-curve analysis suggested that external error had not saturated within the evaluated training range, while sensitivity to input-window length remains untested.

Statistical findings are exploratory because the reported p -values were not adjusted for multiple comparisons. Recording-level bootstrap intervals are conditional on the evaluated trained models and common valid windows, rather than estimates of uncertainty across newly repeated training experiments. The pooled QTDB correlation confidence interval accounted for within-recording clustering through recording-level resampling, whereas the pooled Bland–Altman summaries remain descriptive and do not establish clinical interchangeability. For seed 123, the pooled QTDB correlation of 0.736 contrasted with a repeated-measures correlation of 0.162, indicating substantially weaker within-recording association than the pooled statistic suggests. Several expert-reference analyses used three seeds, further limiting comparability with the four-seed analyses. Uncertainty was not quantified for individual predicted landmarks or QT intervals, and the single-reader clinical reference leaves interobserver variability unmeasured [5].

Deployment requires independent, reference-free end-to-end validation, including cycle detection and explicit accounting for missed or ambiguous beats. QTANet uses

offline bidirectional context and post-processing, and attention increased peak inference memory in the benchmarked implementation. CPU model-inference latency was 200.1 ms for QTANet versus 10.8 ms for the modern U-Net, measured with batch size one, 6-s windows, and eight CPU threads (Table 12). These timings exclude CPU-side wavelet preprocessing, signal acquisition, and the complete measurement pipeline, and should not be interpreted as end-to-end latency. Joint evaluation of predictive performance and computational cost has also been explored in dual-stream Transformer research for visual question answering, including assessments of GPU memory consumption and training time [29]. Further matched-protocol comparisons with contemporary delineators are also needed [19,20]. Priorities include multicenter and multi-annotator validation, genuine prolonged-QT cases, within-patient tracking, predictive uncertainty calibration, and efficiency improvements. The present raw-QT results do not establish clinical readiness or the reliability of autonomous QTc monitoring.

5. Conclusions

QTANet derives QT intervals from explicitly predicted QRS-onset and T-wave-offset landmarks. On the expert-annotated QTDB interval subset, it achieves QT MAE of 34.94 ms with 95.2% coverage, compared with 40.55 ms and 88.2% for Control (– Attention) and 52.41 ms and 68.7% for NeuroKit2. Attention improves expert-reference T-wave F1 on the fully bracketed subset, and a decomposition of false-positive segments shows that, across four seeds, it consistently reduces T waves misread as QRS complexes, with reductions in wave-identity confusion tracking gains in measurement coverage. Against a published U-Net delineator trained under an identical protocol, QTANet is superior in frame-level metrics on both external cohorts and in QT error on QTDB, whereas internally the U-Net has lower QT error, and a larger modern U-Net outperforms QTANet on frame-level metrics and, averaged over four seeds, in QT error on QTDB. A separate window-level analysis reports lower absolute error than the implemented scalar regressor under the evaluated training budget. The evidence is qualified by internal model-selection exposure, differences in reference subsets and seed inclusion, and variability in interval-level results. Reporting error together with coverage and exposing the underlying landmarks provides a useful basis for further investigation of automated QT measurement; clinical utility and deployment reliability remain to be established.

Author Contributions: Conceptualization, S.-B.P. and H.-J.Y.; methodology, S.-B.P. and M.L.; software, S.-B.P.; validation, S.-B.P., M.L. and Y.R.K.; formal analysis, S.-B.P. and M.L.; investigation, M.L. and Y.R.K.; data curation, Y.R.K.; writing—original draft preparation, S.-B.P.; writing—review and editing, M.L., J.H., Y.R.K. and H.-J.Y.; visualization, S.-B.P.; supervision, M.L. and H.-J.Y.; project administration, H.-J.Y.; funding acquisition, H.-J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development program (IITP-2026-RS-2023-00256629) and the AI Star Fellowship Support Program (IITP-2026-RS-2026-25619158), and by the Ministry of Science and ICT (MSIT) under the ITRC support program (IITP-2026-RS-2024-00437718) supervised by the IITP.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Chonnam National University Hospital (protocol code CNUH-2023-167, approved on 16 June 2023).

Informed Consent Statement: Patient consent was waived by the Institutional Review Board because of the retrospective design of the study and the use of de-identified data.

Data Availability Statement: The LUDB and QTDB datasets are publicly available from PhysioNet at <https://physionet.org/content/ludb/> (accessed on 6 October 2026) and <https://physionet.org/content/qtdb/> (accessed on 6 October 2026). The training and evaluation code, the preprocessing pipeline, the trained model weights and the scripts that reproduce the tables and figures are available from the corresponding author on reasonable request. The CNUH dataset is not publicly available owing to privacy and ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, X.; Wang, H.; Li, Z.; Qin, L. Deep learning in ECG diagnosis: A review. *Knowl.-Based Syst.* **2021**, *227*, 107187. [CrossRef]
2. Postema, P.G.; Wilde, A.A.M. The measurement of the QT interval. *Curr. Cardiol. Rev.* **2014**, *10*, 287–294. [CrossRef] [PubMed]
3. Trinkley, K.E.; Page, R.L., II; Lien, H.; Yamanouye, K.; Tisdale, J.E. QT interval prolongation and the risk of torsades de pointes: Essentials for clinicians. *Curr. Med. Res. Opin.* **2013**, *29*, 1719–1726. [CrossRef] [PubMed]
4. International Council for Harmonisation. ICH E14: *The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs*; ICH Harmonised Tripartite Guideline; ICH: Geneva, Switzerland, 2005. Available online: <https://www.ich.org/page/efficacy-guidelines> (accessed on 26 August 2026).
5. Vink, A.S.; Neumann, B.; Lieve, K.V.V.; Sinner, M.F.; Hofman, N.; El Kadi, S.; Schoenmaker, M.H.A.; Slaghekke, H.M.J.; de Jong, J.S.S.G.; Clur, S.-A.B.; et al. Determination and Interpretation of the QT Interval: Comprehensive Analysis of a Large Cohort of Long QT Syndrome Patients and Controls. *Circulation* **2018**, *138*, 2345–2358. [CrossRef] [PubMed]
6. Martínez, J.P.; Almeida, R.; Olmos, S.; Rocha, A.P.; Laguna, P. A wavelet-based ECG delineator: Evaluation on standard databases. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 570–581. [CrossRef] [PubMed]
7. Makowski, D.; Pham, T.; Lau, Z.J.; Brammer, J.C.; Lespinasse, F.; Pham, H.; Schölzel, C.; Chen, S.H.A. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav. Res. Methods* **2021**, *53*, 1689–1696. [CrossRef] [PubMed]
8. Laitala, J.; Jiang, M.; Syrjälä, E.; Naeini, E.K.; Airola, A.; Rahmani, A.M.; Dutt, N.D.; Liljeberg, P. Robust ECG R-peak detection using LSTM. In Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC '20), Brno, Czech Republic, 30 March–3 April 2020; pp. 1104–1111. [CrossRef]
9. Sun, A.; Hong, W.; Li, J.; Mao, J. An arrhythmia classification model based on a CNN-LSTM-SE algorithm. *Sensors* **2024**, *24*, 6306. [CrossRef] [PubMed]
10. Hu, R.; Chen, J.; Zhou, L. A transformer-based deep neural network for arrhythmia detection using continuous ECG signals. *Comput. Biol. Med.* **2022**, *144*, 105325. [CrossRef] [PubMed]
11. Alam, R.; Aguirre, A.D.; Stultz, C.M. QTNet: Deep learning for estimating QT intervals using a single lead ECG. In Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023; pp. 1–4. [CrossRef] [PubMed]
12. Alam, R.; Aguirre, A.D.; Stultz, C.M. Detecting QT prolongation from a single-lead ECG with deep learning. *PLoS Digit. Health* **2024**, *3*, e0000539. [CrossRef] [PubMed]
13. Plagwitz, L.; Doldi, F.; Magerfleisch, J.; Zotov, M.; Bickmann, L.; Heider, D.; Varghese, J.; Eckardt, L.; Büscher, A. QTcNet: A deep learning model for direct heart rate corrected QT interval estimation. *Europace* **2025**, *27*, euaf274. [CrossRef] [PubMed]
14. Kalyakulina, A.; Yusipov, I.; Moskalenko, V.; Nikolskiy, A.; Kozlov, A.; Kosonogov, K.; Zolotykh, N.; Ivanchenko, M. Lobachevsky University Electrocardiography Database (Version 1.0.0). *PhysioNet* **2020**. [CrossRef]
15. Laguna, P.; Mark, R.G.; Goldberger, A.; Moody, G.B. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In Proceedings of the Computers in Cardiology 1997, Lund, Sweden, 7–10 September 1997; Volume 24, pp. 673–676.
16. Joung, C.; Kim, M.; Paik, T.; Kong, S.-H.; Oh, S.-Y.; Jeon, W.K.; Jeon, J.; Hong, J.-S.; Kim, W.-J.; Kook, W.; et al. Deep learning based ECG segmentation for delineation of diverse arrhythmias. *PLoS ONE* **2024**, *19*, e0303178. [CrossRef] [PubMed]
17. Moskalenko, V.; Zolotykh, N.; Osipov, G. Deep learning for ECG segmentation. In *Advances in Neural Computation, Machine Learning, and Cognitive Research III*; Studies in Computational Intelligence; Springer: Cham, Switzerland, 2020; Volume 856, pp. 246–254. [CrossRef]
18. Krasteva, V.; Stoyanov, T.; Schmid, R.; Jekova, I. Delineation of 12-lead ECG representative beats using convolutional encoder-decoders with residual and recurrent connections. *Sensors* **2024**, *24*, 4645. [CrossRef] [PubMed]
19. Park, M.; Lim, J.; Yu, T.; Joo, S. A multi-dataset benchmark for semi-supervised semantic segmentation in ECG delineation. *arXiv* **2025**, arXiv:2507.18323. [CrossRef]
20. Li, R.; Lv, Z.; Wang, M.; Li, Y.-G.; Cai, W. A Mamba-based framework for synchronous delineation of P, QRS, and T wave in 12-lead ECGs. *Biomed. Signal Process. Control* **2026**, *125*, 110774. [CrossRef]

21. Taleban, A.; Sparapani, R.; Noffke, P.; Zlochiver, S.; Lu, Q.; Widlansky, M.E.; Luo, J. Explainable artificial intelligence in electrocardiography: A systematic review. *Biomed. Signal Process. Control* **2026**, *114*, 109325. [[CrossRef](#)] [[PubMed](#)]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Asif, M.S.; Faisal, M.S.; Dar, M.N.; Hamdi, M.; Elmannai, H.; Rizwan, A.; Abbas, M. Hybrid deep learning and discrete wavelet transform-based ECG biometric recognition for arrhythmic patients and healthy controls. *Sensors* **2023**, *23*, 4635. [[CrossRef](#)] [[PubMed](#)]
24. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [[CrossRef](#)] [[PubMed](#)]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; Liu, T.-Y. On layer normalization in the Transformer architecture. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; Volume PMLR 119, pp. 10524–10533.
27. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
28. Bakdash, J.Z.; Marusich, L.R. Repeated measures correlation. *Front. Psychol.* **2017**, *8*, 456, Correction in *Front. Psychol.* **2019**, *10*, 1201. [[CrossRef](#)] [[PubMed](#)]
29. Shehzad, F.; Minutolo, A.; Esposito, M. Designing and evaluating a dual-stream transformer-based architecture for visual question answering. *IEEE Access* **2024**, *12*, 195561–195574. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.