Can I Trust It? A Self-Supervised Approach for Scoring Soft Labels in ECG Segmentation

Jan Pavlus[1,2], Petr Nejedly[1], Hynek Hermansky[2], Gari Clifford[3], Filip Plesinger[1], Jan Cernocky[2]

[1] Institute of Scientific Instruments of the CAS, Brno, Czech Republic
[2] Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[3] Department of Biomedical Informatics, Emory University, Atlanta, GA, United States of America

Abstract

Background: *Automatically generated ECG segmentations enable large-scale analysis, but their reliability varies between records. Estimating segmentation quality is therefore important for deciding which outputs can be trusted.*

Method: *We propose RTrust, a self-supervised reconstruction model trained on reference segmentation masks. Reconstruction loss provides a per-record trust reward, evaluated both as a reference-free confidence estimate and as a weighting or selection criterion for knowledge distillation.*

Results: *RTrust separates reference from degraded masks with AUROC from* 0.566 *to* 0.984 *depending on corruption severity. On model outputs, the in-domain reward correlates with per-record* F_1 *at* $\rho = 0.225$–0.515. *Retaining the highest-reward* 25% *improves pooled macro* F_1 *from* 0.754 *to* 0.796 *on QTDB and from* 0.807 *to* 0.866 *on LUDB. In contrast, reward-guided weighting or data selection changes held-out* F_1 *by less than* 0.006.

Conclusion: *RTrust provides a useful reference-free estimate of ECG segmentation reliability, but does not yield a measurable benefit when used to guide distillation.*

1. Introduction

ECG delineation identifies the P wave, QRS complex and T wave together with their onset and offset boundaries, providing the temporal structure required for clinically relevant waveform and interval measurements [1, 2]. Automatic delineation enables these measurements to be obtained consistently across large ECG collections.

Training segmentation models for this task requires delineated ECGs, but detailed expert annotations remain scarce. Automatically generated segmentation labels offer a practical way to increase the amount of training data, for example through knowledge distillation from a teacher model [3]. Such labels, however, are model predictions rather than expert references, and their reliability can vary between records. The same uncertainty arises when a segmentation model is applied to previously unseen ECGs, where an expert reference is generally unavailable to determine whether an individual prediction should be trusted.

This paper investigates whether a reconstruction-based trust model (RTrust) can provide a reference-free estimate of segmentation reliability. The primary objective is to determine whether the resulting reward reflects the quality of unseen segmentation labels and model outputs, including across different annotation protocols. As a secondary objective, the same reward is used during knowledge distillation to investigate whether emphasizing more reliable teacher-generated labels can improve student training.

2. Methodology

2.1. Data

The *QT Database* (QTDB) [1] is evaluated using three references: the automatic `ecgpuwave` consensus pu and the manual annotations $q1c$ and $q2c$. Signals are divided into 10,s (5000-sample, 500,Hz) chunks, yielding $n = 9450$ chunks with full pu coverage and $n = 5775$ for the augmentation study. QTDB provides two Holter-style signals, not a 12-lead montage: each is placed in the standard slot its channel name implies, and the remaining slots are filled by cycling them, so all 12 channels carry real ECG rather than zeros (worth 0.41 macro F_1 on pu). The other corpora are natively 12-lead. The *Lobachevsky University Database* (LUDB) [4] contains 200 manually delineated 12-lead ECGs and is used as an additional evaluation dataset with a different annotation protocol.

The study also uses the publicly available PTB-XL soft segmentation masks [5, 6], together with 1M HEEDB-MGH, 0.9M HEEDB-Emory and 1M CODE recordings [7, 8], for approximately 2.9M ECGs in total.

Because the QTDB manual annotations cover only limited portions of each record, evaluation is restricted to re-

gions in which an annotator explicitly committed to a delineation. This gives scorable coverage of 94.7% for pu, 48.4% for $q1c$, and 11.9% for $q2c$. The overall teacher–student and RTrust workflow is summarized in Fig. 1.

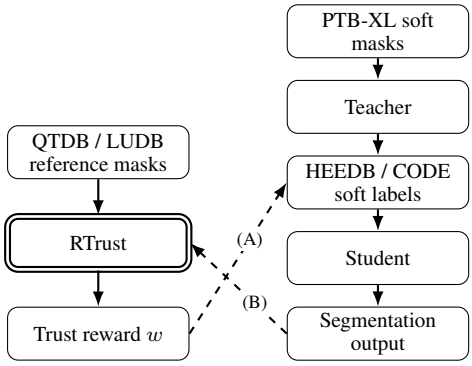

Figure 1. System overview. A teacher generates soft segmentation labels for HEEDB and CODE ECGs without expert delineations, which are used to train a student. RTrust is trained separately on QTDB or LUDB reference masks. The resulting trust reward w (Eq. (1)) is used (A) to weight or select teacher-generated labels during distillation and (B) to estimate the reliability of student outputs.

2.2. Teacher and student networks

The teacher and student use the same 4.6M-parameter U-Net with a three-layer transformer bottleneck ($d_{\text{model}} = 128$), operating on raw 12-lead ECGs and predicting seven segmentation states (unknown, P, PQ, QRS, ST, T, TP).

The teacher is trained on the PTB-XL soft segmentation targets described in Sec. 2.1. It is then applied to the approximately 2.9M HEEDB and CODE ECGs to generate soft segmentation labels for student distillation, as illustrated in Fig. 1. Students are trained from random initialization using knowledge distillation (temperature 2.0, $\alpha = 0.5$), a batch size of 256, and an initial learning rate of 10^{-4}. Training duration is adjusted with the label budget to keep the number of optimization steps approximately comparable.

2.3. Reconstruction-based trust model

RTrust is a self-supervised autoencoder over segmentation masks; it receives the mask alone, never the ECG signal. Two separate models are trained with the same objective: one on QTDB pu reference masks (105 records) and one on LUDB reference masks. Each model reconstructs its input mask under a masked cross-entropy loss ℓ, with lower reconstruction loss indicating greater similarity to the segmentation distribution learned from the corresponding reference corpus, following the broader principle of detecting unexpected inputs through mismatch with learned statistics [9]. Using a reconstruction or likelihood score to decide whether an individual prediction should be trusted is the standard construction in selective prediction and out-of-distribution detection [10, 11].

For scoring, reconstruction loss is converted to a trust reward

$$w = \exp\big(-(\ell - \tilde{\ell})\big), \qquad \tilde{\ell} = \text{median}(\ell), \qquad (1)$$

where $\tilde{\ell}$ is computed for the corresponding RTrust model, so $w = 1$ represents its median reconstruction quality. We refer to w as the reward throughout; it serves directly as the per-record loss weight in Sec. 3.3, hence the weight notation. The QTDB- and LUDB-trained models are both evaluated on model outputs, allowing in-domain and cross-protocol reliability to be compared. Both downstream uses of the reward are summarized in Fig. 1.

2.4. Validation tasks and targets

We evaluate RTrust in three settings. First, *separability* is tested by degrading QTDB reference masks using shifts, boundary perturbations, wave removal/insertion and label swaps at four severity levels, and measuring how well reconstruction loss separates degraded from reference masks. Second, *reliability* is evaluated by scoring the student's own output and comparing the reward with segmentation quality on QTDB pu, $q1c$, $q2c$ and LUDB using Spearman ρ and risk–coverage analysis. Third, *training gain* tests whether the same reward improves distillation, used either as a per-record loss weight or by selecting the top-N teacher-labelled records at a matched label budget. Training contrasts are evaluated with paired bootstrap intervals ($B = 5000$).

Segmentation quality is measured by macro F_1. For risk–coverage curves, per-record confusion matrices are pooled over the retained records before computing macro F_1, avoiding changes in class weighting caused by different class composition across records. Arm-versus-arm training contrasts instead use per-record macro F_1 paired by record, which is well defined because the two arms are

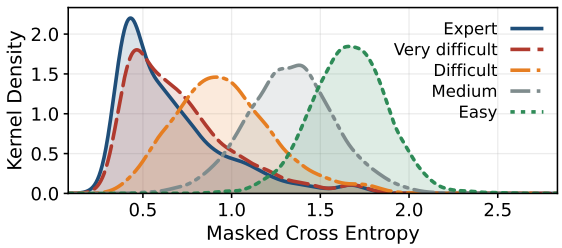

Figure 2. RTrust reconstruction-loss distributions for reference masks and four corruption levels ($n = 5775$ each); lower loss maps to a higher reward in Eq. (1).

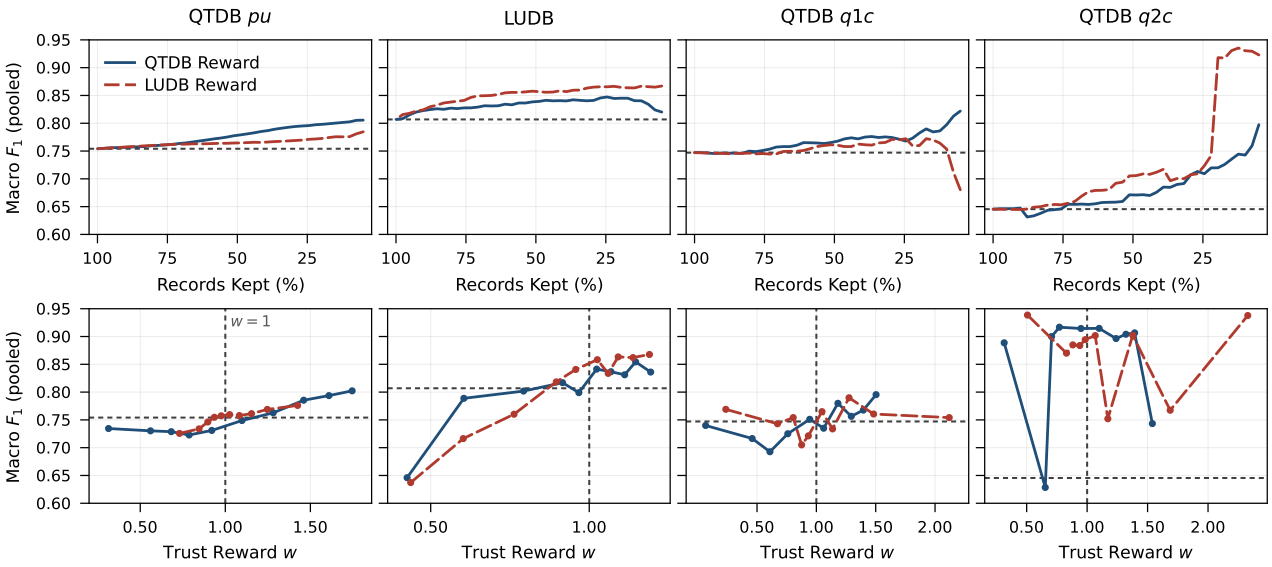

Figure 3. Trust reward against student quality on four evaluation references for the QTDB-trained (blue) and LUDB-trained (red) RTrust models. Both score the student's *own output*. **Top:** risk–coverage curves obtained by retaining progressively higher-reward records; the dashed horizontal line is the full-coverage baseline. **Bottom:** pooled macro F_1 in ten equal-count reward deciles, plotted against the median reward of each decile; $w = 1$ marks the median reconstruction quality of the corresponding RTrust model. Each bottom-row point therefore pools one tenth of the records — only 52 for $q1c$ and 7 for $q2c$, so those two curves are correspondingly noisy. Each RTrust model performs best on its own corpus.

scored on an identical record set. QTDB manual references are scored only over regions in which the annotator provides a valid delineation.

3. Results

3.1. Separation of degraded labels

RTrust increasingly separates reference masks from synthetically degraded masks as corruption severity grows. AUROC is 0.566, 0.794, 0.940 and 0.984 from very difficult to easy corruption, showing strong detection of gross errors but limited sensitivity to subtle boundary perturbations (Fig. 2).

3.2. Reference-free confidence estimation

RTrust is evaluated as a reference-free confidence estimate by scoring the student's own segmentation output and comparing the resulting reward with segmentation quality. Table 1 summarizes ranking and risk–coverage performance. Here, *Base* is pooled macro F_1 using all records, while $\Delta F_1@25\%$ is the change in that metric after retaining only the highest-reward 25% of records.

The reward ranks segmentation quality positively on all four references. The model matched to the evaluation corpus is strongest in each case: the QTDB-trained model reaches $\rho = 0.344$ on pu, 0.225 on $q1c$ and 0.281 on $q2c$ — the latter two being different annotators of the same corpus it was trained on — while the LUDB-trained model reaches $\rho = 0.515$ on LUDB.

Table 1. Reference-free confidence estimation. Base is pooled macro F_1 at full coverage; $\Delta F_1@25\%$ is the change after retaining the highest-reward 25% of records.

		macro	QTDB reward		LUDB reward	
Ref.	n	F_1	ρ	$\Delta F_1@25\%$	ρ	$\Delta F_1@25\%$
pu	9450	0.754	**0.344**	+0.042	0.157	+0.017
LUDB	200	0.807	0.266	+0.040	**0.515**	+0.059
$q1c$	517	0.747	**0.225**	+0.021	0.084	+0.022
$q2c$	68	0.646	**0.281**	+0.065	0.094	+0.066

Cross-protocol ranking stays positive but is asymmetric: the QTDB-trained model reaches $\rho = 0.266$ on LUDB, whereas the LUDB-trained model reaches only 0.084–0.157 on the QTDB references. Transfer is therefore usable from the denser annotation protocol to the sparser one, but weak in the opposite direction.

Selecting only the highest-reward 25% of records improves pooled macro F_1 from 0.754 to 0.796 on QTDB pu, from 0.807 to 0.866 on LUDB, from 0.747 to 0.768 on $q1c$, and from 0.646 to 0.711 on $q2c$. Fig. 3 shows the corresponding risk–coverage and reward–quality curves. Together, these results show that RTrust provides useful record-level ranking without requiring a reference annotation at scoring time.

3.3. Reward-guided distillation

We next tested whether the same reward improves distillation, either as a per-record loss weight or by selecting the highest-reward teacher labels at a matched budget.

Loss weighting. Weighting a per-record loss by a learned reliability score is the usual remedy for label noise in distillation [12, 13]. Here that weight is the reward w of Eq. (1), applied unchanged. Reward weighting produced no transferable gain. On QTDB, changes in held-out F_1 ranged from -0.0055 to $+0.0040$ across the tested budgets, while all LUDB effects remained within ± 0.002. Fig. 4 shows that the median-centred rewards redistribute little training mass: the lowest-weighted half of the records still carries 36% and 42% of the total weight for QTDB and LUDB, respectively, compared with 50% under uniform weighting (Gini 0.19 and 0.11).

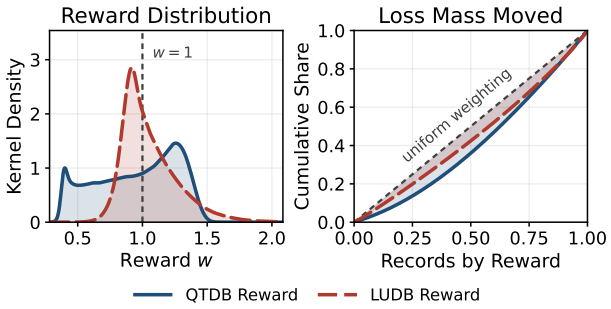

Figure 4. Training rewards (left) and their cumulative share of total loss weight (right). Median-centred rewards move only a small fraction of the training loss.

Data selection. The reward ranking was also used to select the top-N teacher-labelled records instead of a random N at the same label budget. Across the 30k, 100k and 300k budgets and the pu, $q1c$ and LUDB references, all nine contrasts remained within ± 0.006 F_1. LUDB stayed weakly positive ($+0.0012$ to $+0.0039$), whereas the QTDB pu and $q1c$ effects changed sign across budgets. Thus, neither loss weighting nor reward-based data selection produced a consistent training improvement at the scale resolved by these experiments.

4. Conclusion

RTrust turns a segmentation model's own output into a reference-free reliability score: it ranks quality positively on all four references, and keeping the highest-reward quarter of the records raises pooled macro F_1 by 0.02 to 0.07 — useful wherever no expert reference exists. Transfer across protocols stays positive, so a trust model is not confined to the corpus that trained it, though an in-domain model remains stronger. Weighting or selecting distillation labels by the same reward gave no consistent gain; the evidence points to granularity, since one weight per record cannot say which parts of that record are unreliable. Sub-record weighting is the natural next step, alongside the boundary sensitivity, annotation overlap and sample-size limits that bound this analysis.

Acknowledgments

This research was supported by the Czech Technological Agency (grant number FW06010766) and by the Czech Academy of Sciences (RVO68081731).

References

[1] Laguna P, Mark RG, Goldberg A, et al. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. Comput Cardiol 1997; 24:673–676.

[2] Martínez JP, Almeida R, Olmos S, et al. A wavelet-based ECG delineator: evaluation on standard databases. IEEE Trans Biomed Eng 2004;51(4):570–581.

[3] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv150302531 2015;.

[4] Kalyakulina A, Yusipov I, Moskalenko V, et al. Lobachevsky University Electrocardiography Database: a new resource for ECG delineation research. Physiol Meas 2020;41(1):015001.

[5] Pilia N, Nagel C, Lenis G, et al. ECGdeli: an open source ECG delineation toolbox for MATLAB. SoftwareX 2021; 13:100639.

[6] Wagner P, Mehari T, Haverkamp W, et al. PTB-XL (v1.0.1) soft segmentations (delineation). Zenodo, 2023.

[7] Koscova Z, Li Q, Robichaux C, et al. The Harvard–Emory ECG Database. Sci Data 2026;13(1):516.

[8] Ribeiro ALP, Paixão GMM, Gomes PR, et al. Tele-electrocardiography and bigdata: the CODE (Clinical Outcomes in Digital Electrocardiography) study. J Electrocardiol 2019;57:S75–S78.

[9] Hermansky H. Multistream recognition of speech: dealing with unknown unknowns. Proc IEEE 2013;101(5):1076–1088.

[10] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proc ICLR 2017;.

[11] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. Proc NeurIPS 2017;4878–4887.

[12] Jiang L, Zhou Z, Leung T, et al. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. Proc ICML 2018;2304–2313.

[13] Northcutt CG, Jiang L, Chuang IL. Confident learning: estimating uncertainty in dataset labels. J Artif Intell Res 2021; 70:1373–1411.

Address for correspondence:
Jan Pavlus
Institute of Scientific Instruments of the CAS,
Kralovopolska 147
612 00 Brno, Czech Republic
jan.pavlus@isibrno.cz