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

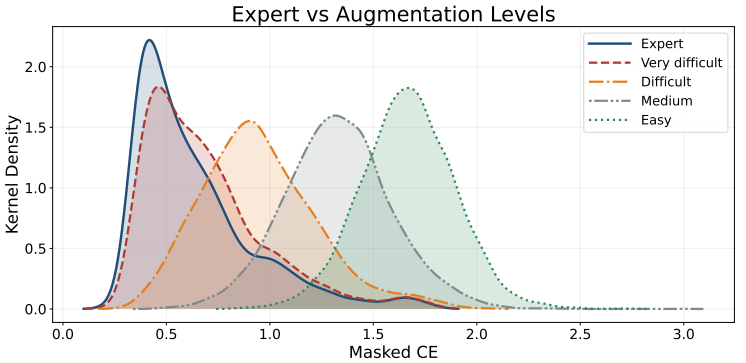
Jan Pavlus, Petr Nejedly, Hynek Hermansky, Filip Plesinger, Jan Cernocky

Institute of Scientific Instruments of the CAS
Brno, Czech Republic
Faculty of Information Technology of Brno University of Technology
Brno, Czech Republic

Background: This work addresses the reliability of teacher-generated soft labels in ECG waveform segmentation. Assessing the quality of such labels is essential for distinguishing between high- and low-quality annotations when ground-truth expert labels are unavailable. The aim of this work is to develop a self-supervised method for scoring soft labels, enabling their use in large-scale datasets. Such a capability is critical for training data-driven models and can further support iterative optimization frameworks, including reinforcement learning-based training.

Method: This work utilizes the QT Database (QTDB) with expert-provided segmentation labels to train a self-supervised reconstruction-based trust model (RTrust). The model learns the distribution of high-quality segmentations by reconstructing reliable examples, with reconstruction fidelity serving as a proxy for trust. During inference, segmentations that are accurately reconstructed are assigned higher trust, whereas those with larger reconstruction errors are considered unreliable. For evaluation, expert annotations are systematically augmented to simulate varying levels of label degradation, including shifts, boundary perturbations, waveform removal, insertion, and label swapping.

Results: Experimental results demonstrate a clear separation of reconstruction error distributions across augmentation levels. Trust-loss distributions differ significantly between expert and augmented segmentations across all levels (Kolmogorov–Smirnov and Mann–Whitney U tests, both $p < 10^{-6}$), with the strongest separation observed for easy and medium perturbations and substantially weaker separation for very difficult cases. This trend is reflected by a decrease in AUROC from 0.984 (easy) to 0.566 (very difficult), accompanied by corresponding reductions in effect size (Cliff's δ from -0.969 to -0.131).



Distribution of reconstruction loss (masked cross-entropy), where lower values indicate better segmentation across categories defined by augmentation level from very difficult (small augmentations) to easy (large augmentations).