Time-efficient labeling framework for biological waveform data using semi-supervised learning and interactive visualization.

Danny Eytan, MD, PhD[1,2], Dmitrii Shubin [1,3], Minfan Zhang[4,5], Daniel Ehrmann, MD[1], Sebastian Goodfellow, PhD [1,3]

[1] Department of Critical Care Medicine, The Hospital for Sick Children, [2] Faculty of Medicine, Technion, Israel,
[3] Faculty of Applied Science and Engineering, University of Toronto, [4] Department of Computer Science, University of Toronto, [5] Vector Institute

Background. The development of machine learning (ML) models for deployment into the clinical setting requires a substantial amount of high-quality labelled data. Generating high-quality labelled datasets for complicated physiologic waveform data, such as electrocardiogram (ECG) data, requires the time of clinical domain experts. Unlike other domains, such as autonomous vehicles where the labelling tasks are more general (draw a box around a cyclist), in many ML for health applications, the labeller needs to be a domain expert. As the time of domain experts is usually scarce, labelling often becomes the rate-limiting step of many ML projects. In this abstract, we propose a novel framework for labelling medical waveform data by using a human-in-the-loop semi-supervised learning pipeline and an interactive visualization approach. We demonstrate the feasibility of the proposed method on labelling ECG waveforms for the classification of a common and deleterious pediatric arrhythmia: Junctional Ectopic Tachycardia (JET).

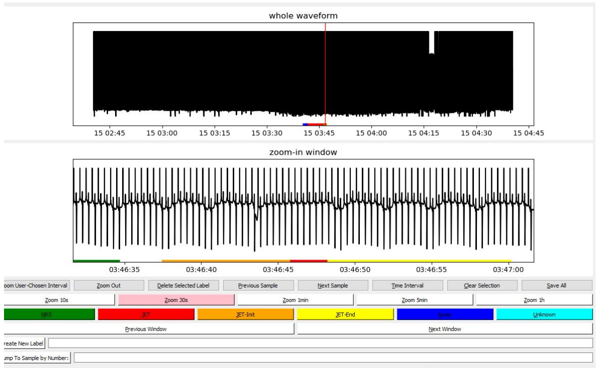

Figure 1. A demo of Waveform Visualization Application

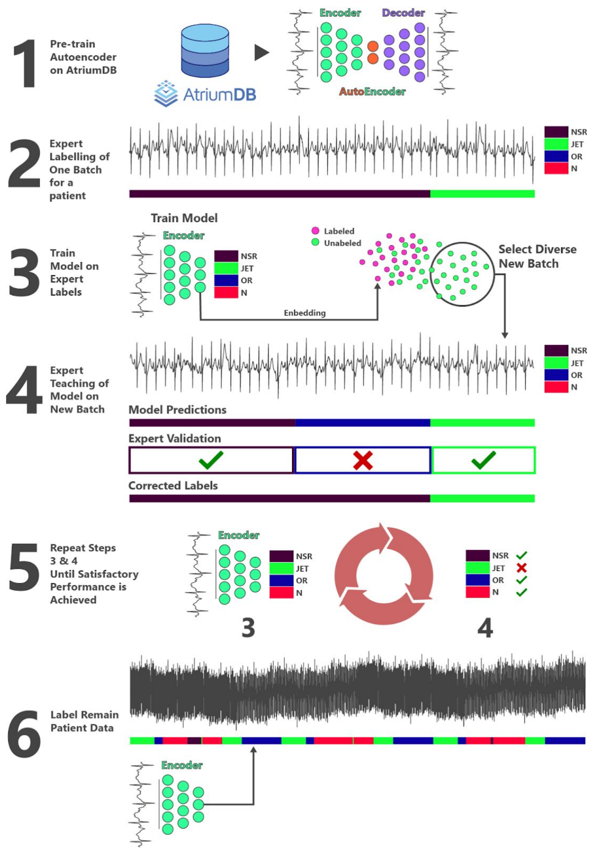

Figure 2. Proposed annotation framework

Methods. Most of the variability of waveform shapes results from patient-wise differences (periodic biological rhythms are mostly repetitive over time), and this property of medical waveforms can be utilized to reduce the amount of data that needs to be annotated manually. The proposed method includes two components: the software tool for biological waveform visualization and a back-end ML engine to assist the labelling process.

The waveform visualization application ((WVA), Figure 1) allows domain experts to assign labels to time-series segments (e.g. JET, normal sinus rhythm, noise, etc.). Additionally, the user can create customized labels as appropriate. The upper window visualizes an overview of a 2-hour waveform block with the ability to zoom in on a chosen segment for labelling efficiency (i.e. if the heart rate suddenly changes). The lower window shows the zoomed-in view of an ECG interval. The timescale can be adjusted to show the waveform with different levels of granularity.

We have prototyped a deep learning-based system to integrate with the WVA by proposing pseudo-labels for the domain expert (Figure 2). We initially trained a deep neural network (DNN) in an unsupervised way using unlabeled data from our hospital's ECG waveform database. The embeddings of the DNN are used to select a diverse segment of a given patient's waveform within the WVA. The domain expert provides labels for this diverse data that is then used to fine-tune the DNN using a one-shot learning approach. Labels can then be proposed to the domain expert for a new segment of diverse waveform data that can be accepted or corrected, thereby further fine-tuning the model. By repeating this process, the model overfits one patient record, which results in more accurate label proposals and fewer segments required for manual labelling after each iteration.

Results. As proof of concept, we implemented the workflow for one iteration (Step 5, Figure 2). Using the WVA, we have labelled ECG data collected from 13 patients (26 patient-hours). We selected 4 patients with a reasonable class balance and extracted feature embeddings by using WaveNet, pre-trained on 6000 hours unlabeled ECG waveforms. Using a KNN model trained on the feature embeddings without fine-tuning, we were able to get pseudo-labels with an 80% F1 macro score on 111 minutes of unseen test data by using only 9 minutes of labelled ECG signals. By implementing the full iterative workflow presented in Figure 2, small batches of diverse samples would continue to be presented to the labeller to further improve the model's performance using as little of the labeller's time as possible.

Conclusions. In this study, we have laid out a framework for ML-assisted waveform labelling, which addresses a major issue in ML for healthcare. Based on the positive results from the proof of concept, we plan to implement the complete pipeline presented in Figure 2. In addition, future research is needed to understand the best ways to mitigate the impacts of mislabelling using this approach and identify the optimal labelling strategy for imbalanced datasets.