Impact of Label Granularity on Photoplethysmography Based Atrial Fibrillation Detection

Yangyang Zhao, Matti Kaisti, Olli Lahdenoja, Tero Koivisto
University of Turku, Turku, Finland

Background: Photoplethysmography (PPG)-based atrial fibrillation (AF) detection has shown promising performance in wearable applications. However, most studies still use a binary AF versus non-AF setting, where normal sinus rhythm and different non-AF arrhythmias are grouped into a single non-AF class. This simplified setup may miss important sources of misclassification, especially in elderly hospitalized patients with cardiovascular comorbidities.

Objective: This study investigates how label granularity affects PPG-based AF detection performance and which clinically relevant categories of non-AF rhythms are more likely to be misclassified as AF.

Methods: Two datasets are used: a private clinical dataset including 49 elderly hospitalized patients and the publicly available MIMIC-III-Ext-PPG dataset. Both datasets contain more than 20 rhythm labels. Within each dataset, the same experimental pipeline is applied while varying only the label definitions, from binary AF versus non-AF classification to finer-grained multiclass settings. A deep learning-based temporal classification model with a residual network and temporal attention is used to compare AF-related performance and confusion patterns across label granularities. Model performance is also evaluated across motion levels using accelerometer-derived motion scores.

Results: Preliminary analysis indicates that increasing label granularity is associated with reduced AF detection performance, mainly due to increased confusion with certain non-AF rhythm categories. AF and normal sinus rhythm remain relatively well separated, whereas performance declines after introducing clinically relevant non-AF rhythms with irregular patterns. Rhythm categories with ectopic or mixed abnormal beats appear more likely to be misclassified as AF. Similar trends are observed across both datasets. Performance also tends to degrade under higher motion levels.

Conclusion: This study examines PPG-based AF detection from the perspective of label design and rhythm composition. These findings may improve understanding of AF misclassification and support evaluation settings that better reflect real clinical conditions.