Towards Trustworthy Atrial Fibrillation Classification from Wearables Data: Quantifying Model Uncertainty

Ciaran A Bench, Nils Strodthoff, Philip J Aston, Andrew J Thompson

National Physical Laboratory
Teddington, United Kingdom

Introduction: Wearable devices capable of measuring Photoplethysmography (PPG) signals are increasingly used to monitor patient health outside of typical clinical settings. PPG signals encode information about relative changes in blood volume and, in principle, can be used to assess various aspects of cardiac health non-invasively, e.g. to detect Atrial Fibrillation (AF). Machine learning based techniques have clear potential to automate diagnostic protocols for AF, where deep networks have been shown particularly effective. However, these models are prone to learning biases and lack interpretability, leaving considerable risk for poor generalisability and misdiagnosis. This makes them unsuitable for routine use in clinical workflows, where the uncertainty/trustworthiness of a model's output is needed to establish whether it can reliably inform diagnoses. Here, we describe the use of Monte Carlo Dropout to estimate the uncertainties of deep learning models trained to predict AF from PPG time series.

Methods: A ResNet-based architecture was trained on raw time series from the DeepBeat dataset to predict AF, achieving performance comparable with the existing literature. During evaluation, the uncertainty for a given prediction is estimated from the distribution of predictions acquired using various forms of sampling. We found that the dropout rate used to parameterise the model has a significant effect on the magnitude of the estimated uncertainties. We propose a grid search to derive rates that produce well-calibrated uncertainties. Furthermore, we formulate a method to disentangle the aleatoric uncertainty (irreducible data uncertainty) from the total estimated uncertainty of each prediction, allowing us to draw insights about the performance of the classifiers.

Results: Fig 1 contains the uncertainty calibration curve (binned estimated uncertainties from the test set vs. average number of incorrect predictions in each bin). The low uncertainty calibration error (UCE) indicates that the estimated uncertainties can be used to assess the trustworthiness of the model's predictions, improving its suitability for use in diagnostic procedures.

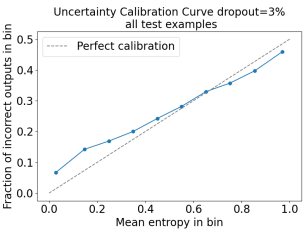

Fig 1: Uncertainty Calibration Curve, (UCE $= 0.033$)