HR Calibration & Forecasting using PPG-Dalia Dataset

# *Intro*

The aim of this project is to utilise the PPG-Dalia Dataset to create a calibration and forecasting model from the same model architecture. Forecasting will be 60 seconds in the future.

# *Raw Data*

PPG-Dalia's data was derived from 15 subjects recorded by ETH Zurich researchers and published via the UCI Machine Learning Repository. Each subject was asked to perform a sequential set of activities for 2.5 hours, for which their ECG-detected HR (via a RespiBAN) and other physiological metrics (via Empatica E4) were measured.

The Empatica E4 measured the following:

- Blood Volume Pulse (BVP)
- Accelerometer (ACC)
- Skin Temperature (TEMP)
- Electrodermal Activity (EDA)

BVP is an estimate of HR, but typically suffers from inaccuracy due to motion artefacts - e.g. certain movements can distort the BVP waveform. Therefore, the presence of ACC, TEMP, and EDA look to counteract this problem by providing additional indicators of the subject's HR.

- ACC gives the model an insight into the individuals' movement at a particular point in time.
- TEMP reports the skin's temperature; higher levels suggest vasodilation which is associated with higher bodily exertion.
- EDA measures the skin's ability to conduct electrical current; higher levels suggest increased sweat gland activity which is associated with higher bodily exertion.

# *Data Processing*

Raw data for each subject was provided without separation, meaning that the data was just one continuous strip of measurements. This format was incompatible with our calibration aim, as we needed a 'section' of Empatica E4 measurements to correspond to an ECG-measured HR*. Therefore, we split the data into 8-second 'windows', for which every window (features) corresponded to an HR (label).

*Note that this ECG-measured HR was an instantaneous mean of the HRs recorded over the corresponding 8-second window.

The Empatica E4 measured the variables at the following rates:

- BVP: 64 Hz
- ACC: 32 Hz

- TEMP: 4 Hz
- EDA: 4 Hz

This disparity meant that for every window, there would be 512 BVP samples compared to merely 32 for EDA and TEMP - going against the model's requirement that all signals be of equal length to ensure time alignment. To resolve this issue, data for ACC, TEMP, and EDA were interpolated to 512 samples.

## *Time Splits*

As per industry standard, the windows were to be randomly split into train, validation, and test sets. However, a split via a simple ratio would not work due to temporal data leakage between sets.

Clarification: splitting the data, for example, as 70:15:15 would mean that some of the windows in the test set (15%) could be chronologically close to windows in the train set (70%). High closeness implies that certain windows from both sets will have very similar feature and label data. This means that rather than predicting some of the test set using learned patterns, the model will simply just 'remember' the corresponding labels of similar training windows, creating the (false) perception that the model is performing well.

To prevent this, the split will take the following composition:

- Train: first 80% of each activity (including transient, ID: 0) from each train subject
- Validation: final 20% of each activity (including transient, ID: 0) from each train subject.
- Test: one of 15 subjects' entire 2.5 hour dataset.

This split takes the Leave-One-Subject-Out (LOSO) approach, for which we train/val/test as many times as there are subjects. Each fold (run) will have a different subject as the test set so we can get a good understanding of the model's generalizability.

Validation taking the final 20% of each activity ensures that the validation's coverage has good breadth, considering that different activities - performed at different times within the 2.5 hour block - induce different HR levels and trends. An embargo will be implemented in the final 8 seconds (1 window) of train data to prevent leakage into validation data.

Windows in train, validation, and test sets that did not have a corresponding t+H ECG-measured HR were removed. These were likely windows towards the end of the 2.5 hour period.

Running tests for all 15 subjects is crucial, considering that the dataset's authors report that subjects' age, gender, weight, and fitness vary.

Train, validation, and test sets were normalised using statistics calculated from the train data.

# *The Model*

The model's architecture consists of a shared encoder (GRU) followed by two heads: one representing calibration and the other forecasting.

A Gated Recurrent Unit (GRU) was used due to the sequential nature of the windows' samples. The model requires an understanding of the relationship between samples for calibration (every sample only representing 1/64 of a second) and forecasting (identification of a trend for future HR prediction). It is the responsibility of the GRU to determine the information in the hidden states that best allows the two heads to derive patterns useful for their respective agendas.

The two heads are simple feed-forward networks. Their functions (calibration and forecasting) are distinguished by their respective target labels.

# *Training*

Model training amongst all 15 datasets saw the same trend: improvements in val_MAE loss until ~ 50 epochs, then overfitting for the remaining epochs. As expected, val_now was typically lower than val_fut (except for fold 15).

- Best val_MAE: 14.7858 [FOLD 13]
- Worst val_MAE: 20.6867 [FOLD 10]
- Average val_MAE: 17.7819 (now: 8.5187, fut: 9.2630)

The strong validation performance when subject 13 was excluded suggests their data was more difficult to model. Subject 10 is likely a more easy dataset to decipher given the low performance in its absence.

Fold 7 saw a weird spike in training loss at ~ epoch 60 despite decreasing up until that point. A theory for its occurrence could be gradient overshooting on outlier windows.

# *Evaluation*

Predictions of test_MAE given val_MAE were inaccurate. Subject 13 did not have the worst performance and subject 10 did not have the best performance. There were no statistical outliers.

- Best test_MAE: 20.4885 (now: 9.7095, fut: 10.7790) [SUBJECT 9]
- Worst test_MAE: 59.0583 (now: 28.3316, fut: 30.7267) [SUBJECT 5]
- Average test_MAE: 31.2515 (std 12.5022) (now: 15.1138 (std 6.31), fut: 16.1377 (std: 6.22))

It can be inferred that subject 9's physiology is most representative of all other subjects, whereas subject 5's is least representative. It should be noted that subject 6 reported the second-worst MAE and has a fitness level of 1, suggesting that her tail-end findings could be due to her lack of fitness.

Important statistics for reference:

- Natural variation between individuals' HR is normal: 20-30bpm when resting, 5-15bpm when submaximal activity, and 10-20bpm for high-intensity activity.

Statistics for current consumer wearables* in market (vs ECG):

- Reported MAE: ~ 7-10bpm
- Reported MAE (rest): ~ 3-5bpm
- Reported MAE (exercise): ~ 10-15bpm

Statistics for current research-grade wearables in market (vs ECG):

- Reported MAE: ~ 3-6bpm

Initially, one would think that an average now_MAE of 15.1138 is passable, however given that nearly half (46.7%) of the data was from high-motion activity, a 15.1138 is a positive result (very close to consumer wearables' ~ 10-15bpm for exercise).

% of time in activities of…

- Rest (low motion, low artefacts): 40 mins → 26.7%
- Moderate (some movement, moderate artefacts): 40 mins → 26.7%
- High (strong artefacts): 23 mins (+ 47 mins from transient) → 46.7%

Forecasting will inevitably be more difficult than calibration, hence the higher average test_MAE. This is due to the randomness of what individuals will be doing 60 seconds from now.

*consumer wearables include but are not limited to: FitBit, Garmin, Apple Watch.
*research-grade wearables include but are not limited to: polar, chest straps, clinical PPG.

## ***Potential Improvements***

Reporting per-activity MAE will confirm whether the high motion activities are indeed producing the higher MAEs relative to the low and moderate activities.

Experimenting with different horizons (e.g. H = 15, 45, 75) will test the limits of the model's forecasting capabilities.

The addition of a Conv1D layer that precedes the shared encoder to attenuate motion artefacts.