

Test Procedure

To allow systems to be comparable, please ensure you follow these instructions carefully:

Challenge 1

To calculate the error in S1 and S2 segmentation, your program must perform the following calculation:

$$\delta_k = \frac{\sum_{i=1}^{\lfloor \frac{N_k}{2} \rfloor} (|RS1_i - TS1_i| + |RS2_i - TS2_i|)}{N_k}$$

$$\delta = \sum_{k=1}^J \delta_k$$

where δ_k means the average distance of the k -th sound clip in a dataset; δ is the total error; N_k is the total number of S1 and S2 in the k -th sound clip; $RS1_i$ ($RS2_i$) indicates the real location of S1(S2) of the i -th heartbeat and $TS1_i$ ($TS2_i$) indicates the calculated location of S1(S2) of the i -th heartbeat. J is the total of all the sound clips in the specific dataset.

We provide a locked Excel spreadsheet for you to evaluate your method on the unlabelled set, in the file:

Challenge2_evaluation_sheet.xlsx

We also provide an example of how to use the file in:

Challenge2_evaluation_sheet_example.xlsx

When your method is as good as you can make it, please submit your best results (training and test) to y.deng.11@ucl.ac.uk

These results should be formatted as CSV files, in the same format as the files **Atraining_normal_seg.csv** and **Btraining_normal_seg.csv** (we will also accept the results embedded in the Excel spreadsheet above).

Challenge 2

To calculate the effectiveness of your classification method, we require three metrics to be calculated from the tp , fp , tn and fn values of your approach. For Challenge 1, we require you to calculate:

- 1) Precision
- 2) Youden's Index
- 3) F-score

For Challenge 2, we require you to calculate:

- 1) Precision
- 2) Youden's Index
- 3) Discriminant Power

Where the necessary equations are:

1. Precision provides us with the positive predictive value – the proportion of samples that belong in category c that are correctly placed in category c .

$$precision = \frac{tp}{tp + fp}$$

2. Youden's Index has traditionally been used to compare diagnostic abilities of two tests, by evaluating the algorithm's ability to avoid failure. In Dataset A, we evaluate the Youden's Index of the Artifact category. In Dataset B we calculate the Youden's Index of problematic heartbeats (Murmur and Extrasystole categories combined).

$$\gamma = sensitivity - (1 - specificity)$$

$$\text{where } sensitivity = \frac{tp}{tp + fn}, \quad specificity = \frac{tn}{fp + tn}.$$

3. F-score (Dataset A only): The F-score is evenly balanced when $\beta = 1$. This metric favours precision when $\beta > 1$ and specificity otherwise. Here we set $\beta = 0.9$ and evaluate the F-score of problematic heartbeats in Dataset A (Murmur and Extra Heart Sound categories combined).

$$F = \frac{(\beta^2 + 1) * precision * sensitivity}{\beta^2 * precision + sensitivity}$$

4. Discriminant Power (Dataset B only) evaluates how well an algorithm distinguishes between positive and negative examples. The algorithm is a poor discriminant if $DP < 1$, limited if $DP < 2$, fair if $DP < 3$, and good in other cases. Here we calculate the Discriminant power of problematic heartbeats (Murmur and Extrasystole categories combined).

$$DP = \frac{\sqrt{3}}{\pi} (\log X + \log Y)$$

where $X = \text{sensitivity} / (1 - \text{sensitivity})$, and $Y = \text{specificity} / (1 - \text{specificity})$

Methods will be ranked using a combination of these three metrics.

We provide a locked Excel spreadsheet for you to test your method's effectiveness on the unlabelled set, in file:

Challenge2_evaluation_sheet.xlsx

We also provide an example of how to use the file in:

Challenge2_evaluation_sheet_example.xlsx

When your method is as good as you can make it, please submit your best results (training and test) to y.deng.11@ucl.ac.uk

These results should be formatted as CSV files, in the following format:

1. The result of each audio in Dataset A should be a 1*4 array, where Columns 1-4 represent Normal, Murmur, Extra Heart Sound and Artifact State respectively.
2. The result of audio in Dataset B should be a 1*3 array, where Columns 1-3 represent Normal, Murmur and Extrastole state respectively.

Each row represents a single audio file. Rows are ordered according to the Excel spreadsheet: evaluation spreadsheet-locked.xlsx.

For each row (audio file), a single 1 should indicate which class the data belongs to, with the other classes marked 0 (e.g. for DataSet A, if audio 1 is classified into

Murmur, then it should be recorded 0,1,0,0) You will be penalised if you try to put more than a single 1 in a row.

(We will also accept the results embedded in the Excel spreadsheet above.)

Please note that we also require the code for the method, which needs to include instructions for executing the system, to enable us to validate the submitted results if necessary.