Replication

Results are shown in three places, Table 1, Figure 1 and Figure 2.

Note: Due to the random nature of the data shuffling and splitting, and that dropout is used, there will be variance within the results.

Table 1

Raw and Average Results

The CNN results found in Table 1 can be replicated by running the following commands

- **`python cnn_classification.py --dataset tensorflow`**
- **`python cnn_classification.py --dataset pytorch`**
- **`python cnn_classification.py --dataset keras`**
- **`python cnn_classification.py --dataset incubator-mxnet`**
- **`python cnn_classification.py --dataset caffe`**

The raw results are saved to **`results_cnn/hyperband/{dataset}.csv`**.
The metric averages are printed and are also saved to **`results_cnn/hyperband/averages.csv`**.

The Baseline results found in Table 1 can be replicated by running the baseline solution with the following command. For each run, the code must be changed to change the 'project' variable on line 86.

- **`python br_classification.py`**

The raw results are saved to **`results_baseline/{dataset}.csv`**.
The metric averages are printed and are also saved to **`results_baseline/averages.csv`**.

Statistical Differences

The statistical difference results found in Table 1 can be replicated by running the following command

- **`python statistical_test.py results_cnn/hyperband results_baseline`**

Raw p-values and the confirmation of whether the null hypothesis has been rejected (indicating a statistically significant difference) are printed to the terminal.

The p-value results are also saved to **`statistical_tests/cnn_hyperband_vs_baseline/pvalues.csv`**
The significant difference results are also saved to
`statistical_tests/cnn_hyperband_vs_baseline/significant_differences.csv`

Figure 1

Raw and Average Results

The Baseline results found in Figure 1 uses the Accuracy, Precision, Recall, F1-Score and AUC results from Table 1.

The CNN + Hyperband Tuning results found in Figure 1 uses the Accuracy, Precision, Recall, F1-Score and AUC results from Table 1.

The CNN + Manual Tuning results found in Figure 1 can be replicated by running the following command

- **`python cnn_classification.py --dataset tensorflow --manual-tuned-model`**

The raw results are saved to **`results_cnn/manual/tensorflow.csv`**.
The metric averages are printed and are also saved to **`results_csnn/manual/averages.csv`**.

Statistical Differences

Baseline vs CNN + Hyperband

The statistical differences between the Baseline and CNN + Hyperband Tuning uses the statistical differences from Table 1.

CNN + Hyperband Tuning vs CNN + Manual Tuning

The statistical differences between the CNN + Hyperband and CNN + Manual can be replicated by running the following command

- **`python statistical_test.py results_cnn/hyperband results_cnn/manual`**

Raw p-values and the confirmation of whether the null hypothesis has been rejected (indicating a statistically significant difference) are printed to the terminal.

The p-value results are also saved to **`statistical_tests/cnn_hyperband_vs_cnn_manual/pvalues.csv`**
The significant difference results are also saved to
`statistical_tests/cnn_hyperband_vs_cnn_manual/significant_differences.csv`

Baseline vs CNN + Manual Tuning

Whilst there logically is a statistically significant difference between the Baseline and CNN + Manual Tuning, this can be verified by running the following command

- **`python statistical_test.py results_baseline results_cnn/manual`**

Raw p-values and the confirmation of whether the null hypothesis has been rejected (indicating a statistically significant difference) are printed to the terminal.

The p-value results are also saved to **`statistical_tests/baseline_vs_cnn_manual/pvalues.csv`**
The significant difference results are also saved to
`statistical_tests/baseline_vs_cnn_manual/significant_differences.csv`

Figure 2

The results found in Figure 2 uses the AUC and F1-Score metric results from Table 1 over the different datasets.