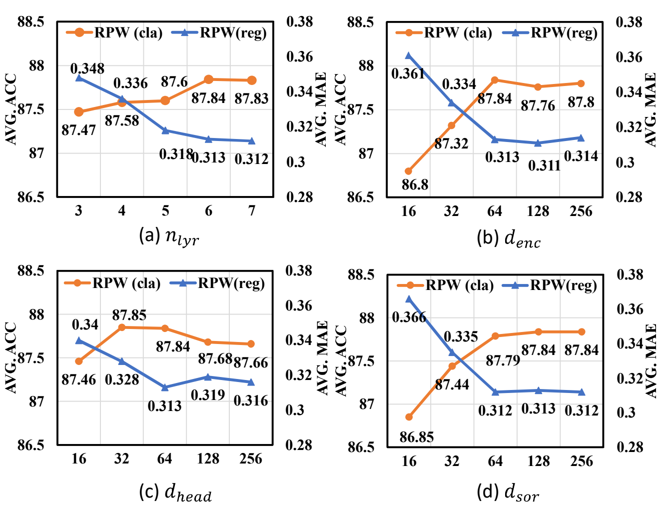
7 HYPER-PARAMETER SENSITIVITY TEST OF SMART

In this section, we test the sensitivity of four main hyper-parameters: i) the number of layers in the encoder n_{lyr}; ii) model dimension d_{enc} (i.e., dimension of the representation for each time point) of the encoder; iii) dimension of head embedding d_{head} in the cross attention; and iv) model dimension of source selector d_{sor}. An overall good setting of these four parameters is: the number of layers in the encoder is 6, the model dimension of the encoder is 64, the dimension of head embeddings in the source selector is 64, and the model dimension of the source selector is 128. To validate the sensitivity of the model performance to these parameters, we vary the parameters by: i) n_{lyr} as 3, 4, 5, 6, and 7; ii) d_{enc} as 16, 32, 64, 128, and 256; iii) d_{head} as 16, 32, 64, 128, and 256; iv) d_{sor} as 16, 32, 64, 128, and 256. Figure 9 shows the results of parameter sensitivity tests. As shown in Figure 9 (a), the classification accuracy increases and the average MAE decreases with more layers, while both converging after $n_{lyr} = 6$. In Figure 9 (b), we observe that setting $d_{enc} = 64$ achieves slightly higher accuracy for classification but a slightly larger MAE for regression than setting $d_{enc} = 128$. To cut down the number of parameters, we believe $d_{enc} = 64$ is a better choice. In Figure 9 (c), we observe that setting d_{head} achieves the second highest accuracy (i.e., 87.84% v.s. the highest accuracy 87.85%) and the smallest MAE when d_{head} is 64. Besides, increasing d_{head} from 64 to a larger number (e.g., 128 and 256) will slightly decrease both the accuracy and MAE, which may be due to too many model parameters, making the model hard to train. Finally, from Figure 9 (d), we observe that the performance of classification and regression converges after the model dimension is 128 or higher. Besides, we observe that when d_{sor} is smaller than 32, the performance is worse, which proves that the model performance is sensitive to the model dimension selected in the source selector and it is necessary to select a large enough d_{sor} (at least 128).



(a) n_{lyr} (b) d_{enc} (c) d_{head} (d) d_{sor}

Figure 9: Parameter Sensitivity Test

8 DETAILS OF VALIDATION AND TESTING GROUPS

In our experiments, we have the following 10 datasets in the validation groups: Computers, RefrigerationDevices, SmallKitchenAppliances, ECG5000, NonInvasiveFetalECGThorax2, MiddlePhalanxTW, SwedishLeaf, CricketZ, InlineSkate, and UWaveGestureLibraryX.

We have the following 10 datasets in the testing groups: WormsTwoClass, ItalyPowerDemand, Lightning2, StarLightCurves, Mallat, Beef, Ham, Meat, InsectWingbeatSound and Strawberry. The other 65 bake-off datasets are in the training groups.

9 BASELINES

We list the baselines evaluated against SMart as follows.

RandOm Convolutional KErnel Transform (ROCKET) [4]: an ensemble which combines models with convolutional filters of random sizes for classification. To facilitate time series regression, we replace the last layer with a linear layer in ROCKET for regression.

Rel-CNN [16]: a time series classifier exploiting relationship features. We also adapt Rel-CNN for regression with a linear layer.

TimeNet [28]: a multi-layered recurrent neural network trained in an unsupervised manner to extract time series representations.

TS-TCC [9]: an unsupervised time series representation learning framework via temporal and contextual contrasting.

TS2Vec [39]: a framework for learning time series representations in arbitrary semantic levels.

DTW-FCN [11]: a fully convolutional network (FCN) based representation transfer framework which selects a single source dataset with the shortest DTW distance to the target dataset to help the representation learning.

TST [40]: a transformer-based representation framework, which learns time series representations by enforcing the representations to recover the masked part of the input time series.