Dataset	cEnsemble	DTW
ScreenType	**0,02**	0,01
UMD	**0,3**	0,27
TwoLeadECG	**0,88**	0
ECG200	**0,32**	0,08
Computers	**0,09**	0
Coffee	**0,8**	0,01
GunPoint	**0,56**	0
Arrowhead	**0,28**	0,2
ItalyPowerDemand	**0,57**	0
Meat	**0,4**	0
OliveOil	0,15	**0,32**
Plaid	**0,35**	0,01
Asphalt Regularity	**0,54**	0
Asphalt Obstacles	**0,38**	0,27
Gesture Pebble	**0,25**	0,02

Table 6: DTW Ablation Test Results.

5.4 Ablation tests for the FeatTS pipeline

In this section, we present a handful of ablation tests devoted to show the importance of each step of the pipeline.

DTW versus Features. The idea behind this ablation test is to show the importance of adopting the distance between the features instead of using the distance between the raw time series using Dynamic Time Warping (DTW) [**?**]. In a nutshell, this corresponds to remove Step 1 and 2 from the pipeline in Figure 2, i.e. the extraction and the selection of the features using PFA.

Table 6 shows the results obtained by replacing the distances between the features, as shown in Figure 4a with the distance computed by the DTW directly on the time series.

The obtained results showed that, 14 datasets (out of 15) have a worse behavior if DTW is adopted. Indeed, only in one dataset (*OliveOil*) the usage of DTW outperforms the usage of the features. On average, we have a difference in terms of AMI of 0.31. Hence, this ablation test confirms the importance of using distances among features instead of merely using distances among raw data.

P-value versus Random. Table 7 shows the results removing the ordering of the features based on their relevance. We repeated this test 5 times and we averaged the results.

The column Rand represents the average of the results obtained by cEnsemble using Random Features.

The results in Table 7 show that the performance of the algorithm drastically deteriorates for the majority of the datasets (13 out of a total of 15 datasets) if random features are employed. Indeed, computing the average overall the results obtained by the two experiments, we have a difference of 0.35 in terms of AMI.

Therefore, the ordering of the features based on their relevance turns out to be indispensable in order to achieve good results.

Dataset	cEnsemble	Rand
ScreenType	**0,02**	0
UMD	**0,3**	0,01
TwoLeadECG	**0,88**	0,02
ECG200	**0,32**	0,06
Computers	**0,09**	0
Coffee	**0,8**	0,12
GunPoint	**0,56**	0
Arrowhead	**0,28**	0,02
ItalyPowerDemand	**0,57**	0
Meat	**0,4**	0,16
OliveOil	0,15	**0,2**
Plaid	**0,35**	0,11
Asphalt Regularity	**0,54**	0
Asphalt Obstacles	**0,38**	0,02
Gesture Pebble	**0,25**	0

Table 7: Random Features Ablation Test

kMeans versus cEnsemble. The purpose of this ablation test is to show the importance of capturing the global relationship among the raw time series samples through graph encoding and of the subsequent application of the Community Detection algorithm as in our approach. Therefore, we replace Step 3,4 and 5 as in Figure 2 with k-Means, i.e. a classical clustering algorithm in its multidimensional version. Hence, once the features have been extracted and selected through PFA in Step 1 and 2, we apply the k-Means algorithm to obtain the clustering.

In Table 8, we show the results obtained capturing the global relationship among the raw time series samples through graph encoding and of the subsequent application of the Community Detection.

The column k-Means shows the results obtained by k-Means among the features selected by each dataset used in our experiment. We have highlighted in bold the best results obtained between k-Means and cEnsemble for each dataset.

The results show that cEnsemble outperforms k-Means in the majority of the cases. Indeed, there are only four datasets for which k-Means shows slightly better results, namely UMD, $Meat$, $Coffee$ and $OliveOil$. On average, the results obtained by cEnsemble outperforms the results obtained by k-Means of 0,08 in terms of AMI.

This ablation test, therefore, shows the importance of capturing the global relationships between the various time series in order to achieve better results in terms of performance.

Dataset	cEnsemble	k-Means
ScreenType	**0,02**	0,01
UMD	0,33	**0,42**
TwoLeadECG	**0,88**	0,75
ECG200	**0,33**	0,16
Computers	**0,09**	0
Coffee	0,89	**0,9**
GunPoint	**0,56**	0,26
ArrowHead	**0,3**	0,24
ItalyPowerDemand	**0,57**	0,51
Meat	0,42	**0,48**
OliveOil	0,18	**0,35**
Plaid	**0,35**	0,03
Ashpalt Regularity	**0,54**	0,35
Asphalt Obstacles	**0,38**	0,37
Gesture Pebble	**0,25**	0,16

Table 8: k-Means Ablation Test results