Technical Report

Time2Feat: **Evaluation**

Angela Bonifati
Lyon 1 Univ., Liris CNRS
angela.bonifati@univ-lyon1.fr

Francesco Del Buono
Univ. Modena e Reggio Emilia
francesco.delbuono@unimore.it

Francesco Guerra
Univ. Modena e Reggio Emilia
francesco.guerra@unimore.it

Donato Tiano
Lyon 1 Univ., Liris CNRS
donato.tiano@univ-lyon1.fr

1 EXPERIMENTAL EVALUATION

Baselines. We selected 18 benchmark datasets from the UEA multivariate time series classification archive [1]. For each dataset, Table 1 reports the number of MTS (V), the number of signals (S), the length (N) of the series, and the clusters (C), where the MTS can be grouped according to the baselines. In addition, we computed the overall number of elements in the dataset (E_O – obtained by multiplying $V \times S \times N$) that provides a yardstick for measuring the scalability of the approach. Finally, we estimate the complexity of generating the clusters by computing the number of elements per MTS (E_M – obtained by multiplying $S \times N$). Intuitively, the lower the value, the lower the ability to extract descriptive features. The datasets represent different scenarios as their overall number of elements E_O spans over three orders of magnitudes, and E_M ranges from 16 elements for the PD dataset to 10000 for SW.

We compared Time2Feat with eight approaches: Hierarchical, KMeans, and Spectral are straightforward applications of these classical clustering techniques to MTS datasets. CPSCA [5] and MC_2PCA [4] introduce a PCA-based mechanism to reduce the data dimensionality before the clustering. DETSEC [2], Reservoir [9] and Dpsom [6] and IT-TSC [10] leverage on neural networks. Finally, we varied the KMeans clustering technique by introducing DTW [7] to measure the similarity between two temporal sequences.

Setup. The experiments are executed on an Intel Xeon Processor machine with 12 cores, 64GB of RAM, and 324GB of local (SSD) storage. The machine runs Ubuntu version 18.04. All experiments have been executed ten times, and the average result plus standard deviation is reported (whenever significant).

1.1 Effectiveness

We evaluated the effectiveness of Time2Feat by adopting the AMI [8] to measure the accuracy of the generated clusters with respect to the baselines. The AMI, takes a value of 1 when the two clusterings are identical, and around 0 (negative values are allowed) in case of random partitions. Table 2 shows the results of this experiment. Time2Feat has been evaluated executing the *unsupervised mode* (column $\mathbf{T2F_0}$) and by simulating the *semi-supervised mode* through stratified random samples composed of 20% (column $\mathbf{T2F_2}$), 40% (column $\mathbf{T2F_4}$), 50% (column $\mathbf{T2F_5}$) of labels per cluster from the baseline datasets. The remaining columns show the competing approaches. Among them, Hierarchical, KMeans, and Spectral can be considered as reference baselines for their simplicity[1]. Finally, in Table 2, we mark in bold the best value per dataset, and with ↑ the results where the selected Time2Feat configuration overcomes the competing approaches while not obtaining the best accuracy value. We do not consider the confidence intervals due to the high discrepancy between the computed values.

[1]We rely on the `sklearn` implementations of these algorithms with default parameters.

Table 1: The datasets evaluated in the experiments. V is the number of MTS, S the number of signals, N the length of the series, C the number of classes in the ground truth, E_O the overall number of elements per dataset, E_M the number of elements per MTS.

Dataset	V	S	N	C	E_O $(VxSxN)$	E_M (SxN)
Li — Libras	360	2	45	15	32400	90
AF – AtrialFibrillation	30	2	640	3	38400	1280
BM – BasicMotions	80	6	100	4	48000	600
RS – RacketSports	303	6	30	4	54540	180
ER – ERing	300	4	65	6	78000	260
Ep – Epilepsy	275	3	206	4	169950	618
PD – PenDigits	10992	2	8	10	175872	16
SW – StandWalkJump	27	4	2500	3	270000	10000
UW – UWaveGestureLibrary	440	3	315	8	415800	945
Ha – Handwriting	1000	3	152	26	456000	456
AW – ArticularyWordRecognition	575	9	144	25	745200	1296
HM – HandMovementDirection	234	10	400	4	936000	4000
LS – LSST	4925	6	36	14	1063800	216
Cr – Cricket	180	6	1197	12	1292760	718
EC – EthanolConcentration	524	3	1751	4	2752572	5253
S1 – SelfRegulationSCP1	561	6	896	2	3015936	5376
S2 – SelfRegulationSCP2	380	7	1152	2	3064320	8064
PS – PhonemeSpectra	6668	11	217	39	15916516	2387

1.2 Interpretability

We provide a measure of the interpretability of the clusters by analyzing the number of features that Time2Feat uses for their computation. A limited number of features helps human comprehension and conciseness is one of the main properties for interpretable features, used in many approaches. The column All in Table 3 shows the overall amount of features extracted after the feature extraction step of the pipeline. The other columns report the number of features retained with the unsupervised mode (column $\mathbf{T2F_0}$) and with increasing levels of supervision as in the previous experiment. The values represent the average of the features selected in ten repetitions of the experiments.

1.3 Efficiency

We perform three experiments to evaluate the efficiency of our approach. The first experiment, in Section 1.3, computes the overall time required to complete the pipeline. The second experiment, in Section 1.3.3, evaluates the time breakdown of Time2Feat's pipeline. Finally, the third experiment, in Section 1.3.4, introduces a simple heuristic to optimize the parallelism of the feature extraction.

1.3.1 Time Performance. Table 4 shows the maximum time to complete the cluster computations for all the datasets in the 10 repetitions of the experiment. We show only the time measured in the unsupervised mode ($\mathbf{T2F_0}$): the semi-supervision does not change the value significantly. The last row shows the average time computed on all datasets (excluding the ones raising the exceptions).

Table 2: Effectiveness (AMI). In bold, the best value per dataset. ↑ shows Time2Feat **settings overcoming all competing approaches.**

Dataset	Semi-supervised			Unsupervised	Competing approaches									
	$T2F_2$	$T2F_4$	$T2F_5$	$T2F_0$	Hier.	KM.	Spe.	DTW	CPSCA	DETSEC	MC_2PCA	*Res.*	*Dpsom*	IT-TSC
Li	0.728±0.02↑	0.722 ± 0.016↑	**0.730±0.020**	0.716±0.012↑	0.563	0.545	0.492	0.503	0.311	0.416	0.069	0	0	0.483
AF	0.028±0.046	0.123 ± 0.066↑	**0.238±0.059**	0.038±0.027↑	-0.002	-0.002	-0.002	0.005	-0.07	-0.001	-0.056	0	0.04	-0.06
BM	0.977±0.034↑	**1.000 ± 0.000**	**1.000±0.000**	**1.000±0.000**	0.347	0.23	0.002	0.832	0.7	**1.000**	0.189	0.57	0.002	0.676
RS	0.559±0.038↑	0.666 ± 0.049↑	**0.710±0.047**	0.35±0.006	0.192	0.194	0.0	0.215	0.221	0.224	0.094	0.13	0	0.41
ER	0.801±0.016	0.823 ± 0.014	0.826±0.023	**0.921±0.011**	0.859	0.91	0.0	0.775	0.5	0.646	0.115	0	0	0.315
Ep	0.896±0.025↑	**0.913 ± 0.019**	0.882±0.007↑	0.792±0.04↑	0.135	0.167	-0.001	0.25	0.258	0.213	0.08	0.12	0	0.68
PD	0.752±0.022↑	0.771 ± 0.028↑	**0.784±0.013**	0.437±0.02	0.728	0.682	N/A	0.6	N/A	0.431	0.065	0.3	0	0.716
SW	0.038±0.036	0.101 ± 0.01	**0.23±0.046**	0.048±0.079	0.131	-0.002	0.0	-0.005	-0.072	-0.097	0.045	0	0	-0.02
UW	0.555±0.026	0.554 ± 0.036	0.59±0.035	0.587±0.055	**0.752**	0.712	0.0	0.611	0.236	0.414	0.111	0	0	0.749
Ha	0.325±0.023↑	**0.353 ± 0.019**	0.349±0.009↑	0.161±0.006	0.226	0.193	0.0	0.235	0.165	0.271	-0.004	0	0	N/A
AW	0.921±0.007	0.931 ± 0.01↑	0.927±0.005↑	**0.963±0.007**	0.926	0.902	0.0	0.781	0.716	0.794	0.182	0	0	0.752
HM	0.021±0.011↑	**0.045 ± 0.007**	0.07±0.012	0.015±0.008	-0.006	-0.002	0.001	0.01	0.002	-0.004	0.018	0	0	-0.01
LS	0.293±0.013↑	0.317 ± 0.011↑	**0.333±0.002**	0.156±0.009↑	0.028	0.018	0.001	N/A	0.047	0.152	0.048	0.09	0	N/A
Cr	**0.984±0.021**	0.975 ± 0.018↑	0.974±0.021↑	0.946±0.021↑	0.756	0.719	0.0	N/A	0.876	0.865	0.361	0	0	0.274
EC	0.065±0.017↑	0.097 ± 0.006↑	**0.121±0.04**	0.052±0.002↑	0.009	0.01	-0.003	N/A	0.013	N/A	0.002	0.03	0	0
S1	**0.397±0.047**	0.374 ± 0.025↑	0.382±0.012↑	0.007±0.001	0.212	0.194	-0.001	N/A	N/A	0.18	0.022	0.1	0	0.08
S2	0.008±0.003↑	**0.015 ± 0.004**	**0.015±0.006**	0.003±0.001 ↑	-0.002	-0.001	0.01	N/A	0.001	0.007	0.005	0.001	0	0.002
PS	0.2±0.006↑	**0.202 ± 0.002**	0.201±0.002↑	0.121±0.007↑	0.093	0.096	0.0	N/A	N/A	N/A	0.058	0.07	0	N/A

Table 3: Number of intra-signal / inter-signal features.

Dataset	All	$T2F_0$	$T2F_2$	$T2F_4$	$T2F_5$
Li	1574/8	55/1	7.17/0	8.33/0	9.4/0
AF	1574/8	21/0	2.83/0.17	5.67/0	5.67/0
BM	4722/120	44.33/1.67	2.33/1.5	2.0/0.67	2/0.17
RS	4722/120	141.2/9.8	12.8/2.6	15.4/3.4	21/4.2
ER	3148/48	125.83/3.17	7/1.67	6.83/1.33	7.17/1.17
Ep	2361/24	163.33/3.67	12.67/1.83	15.67/1.17	15.33/1.33
PD	1574/8	98/1	16.4/0.6	13.8/0.6	18.8/0.8
SW	3148/48	20/0	1.8/0.4	3.4/0	6.4/0
UW	2361/24	124/3	4.4/0.2	4.4/0.2	4.4/0
Ha	2361/24	309.83/3.17	23.83/1.5	25/2.17	30.83/2.67
AW	7083/288	283.67/30.33	10/5	9.5/4	10.5/4
HM	7870/360	167/11	15.17/1	28.17/1.33	24.83/2.17
LS	4722/120	217/5	6.6/3.6	9.4/3.2	12.8/4.4
Cr	4722/120	113.17/3.83	4.5/3.83	4.17/4.17	4.17/4
EC	2361/24	122.83/5.17	9.33/0.33	8.5/0	3/0
S1	4722/120	222.2/1.8	2/0.2	2/0	3/0.2
S2	5509/168	183/3	26.4/0.4	20.8/0.4	20.2/0
PS	8657/440	293/8	4/0	4.4/0	4.2/0

Table 4: Runtime execution , in seconds (- timeout exception fixed in 10 hours, × memory exception).

Dst.	$T2F_0$	Hier.	KMeans	Spec.	DTW	CPSCA	MC_2PCA	DETSEC	IT-TSC
Li	20.31	0.2	0.28	0.37	366	2	0.53	500	31
AF	31.01	0.04	0.06	0.31	350	0.01	0.12	876	57
BM	58.45	0.09	0.16	0.23	175	0.03	209	260	31
RS	50.11	0.2	0.39	0.39	474	0.317	356	270	60
ER	34.2	0.18	0.24	5.27	914	0.21	559	555	85
Ep	47.95	0.24	0.42	7.71	3667	0.31	682	1673	173
PD	198.03	9.0	3.0	×	24713	50	6	3395	7410
SW	559.69	0.16	0.25	0.34	10768	0.01	1	10142	255
UW	58.42	0.45	0.74	15.6	20639	0.47	3063	4229	600
Ha	44.74	0.86	2.0	7.19	27018	0.19	25	4301	-
AW	135.18	1.0	1.0	1.35	23611	1	18811	220	57
HM	220.78	0.83	1.0	0.89	30251	0.62	3162	3175	783
LS	300.23	6.0	3.0	6.49	-	0.35	16591	4666	-
Cr	737.61	0.8	1.0	0.91	-	0.14	13642	12145	2478
EC	876.3	2.0	2.0	2.01	-	0.26	9708	-	2865
S1	727.37	2.0	2.0	2.37	-	-	12	19317	1831
S2	952.58	2.0	2.0	2.22	-	0.36	11	20898	1831
PS	1219.88	1.0	1.0	34.73	-	×	×	-	-
Avg	348.49	1.50	1.14	5.19	11912.17	3.52	3931.69	5537.88	1390.73

1.3.2 Scalability. To evaluate the scalability, we used 27 synthetic datasets generated by varying the number of MTS $V \in$ (100, 1000, 2500, the number of signals $S \in$ (2, 8, 16) and the length of the series $N \in$ (100, 1000, 2000) by means of the open-source API GRATIS [3] 1000, 10000}. The results of the experiments are shown in Figure 1. The setting $V = 2500, S = 16, N = 2500$ generates a memory error in the server used for the experiments.

1.3.3 Time breakdown of the pipeline components. The goal of this experiment is to analyze the breakdown of the computation time into the main pipeline components (feature extraction, feature selection, and cluster generation). The results of the experiments are shown in Figure 2a (logarithmic scale) and in Figure 2b (linear scale).

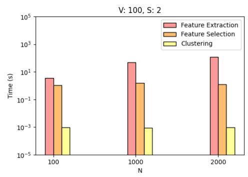
1.3.4 Workload balancing. In this experiment, we evaluate a straightforward heuristic to improve the time performance by optimizing the computational workload on the processors. We recall that feature extraction performs the computation using batches of time series. These batches are not balanced by default. Time2Feat allows to balance the workload by customizing the number of batches per dataset by dividing the total number of MTS (V) by the number of available processors, rounding for excess to the upper integer. Figure 2c (logarithmic scale) and Figure 2d (linear scale) shows the time reduction by adopting this heuristic.

1.4 Robustness

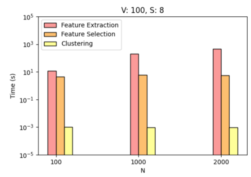
This Section evaluates the robustness of the pipeline components by a series of ablation tests. We evaluate the importance of feature selection (Section 1.4.1). Then, we evaluate alternative options to the Hierarchical algorithm for performing the final cluster computations (Section 1.4.2). Finally, we evaluate the importance of the features in the clustering procedure (Section 1.4.3).

1.4.1 Importance of feature selection. We show the importance of feature selection by evaluating how the accuracy (AMI) in Figure 3a, the interpretability (number of features) in Figure 3b, and efficiency (time for performing the feature extraction and clustering) in Figure 3c vary without any feature selection.

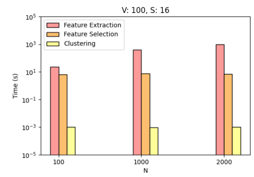
1.4.2 Importance of the clustering technique. We experimented with 3 techniques (*Hierarchical*, *KMeans*, *Spectral*) for generating the clusters, as shown in Table 5. For each technique, we computed the AMI of the clusters obtained with three settings: the *unsupervised procedure* (**$T2F_0$**), the *semi-supervised procedure* with 20% and 50% labeled elements per cluster (**$T2F_2$** and **$T2F_5$**, respectively).



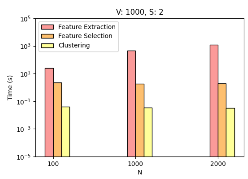
(a) Setting V=100, S=2



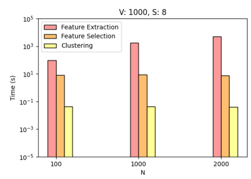
(b) Setting V=100, S=8



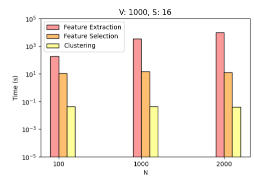
(c) Setting V=100, S=16



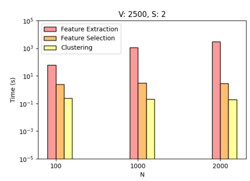
(d) Setting V=1000, S=2



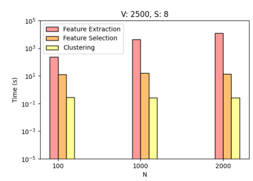
(e) Setting V=1000, S=8



(f) Setting V=1000, S=16

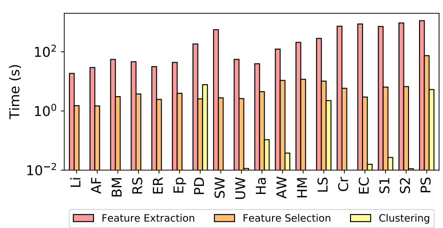


(g) Setting V=2500, S=2

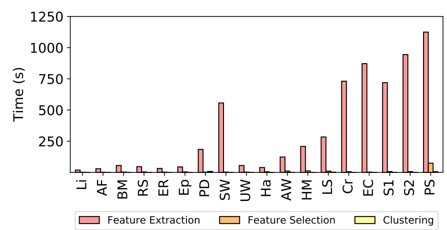


(h) Setting V=2500, S=8

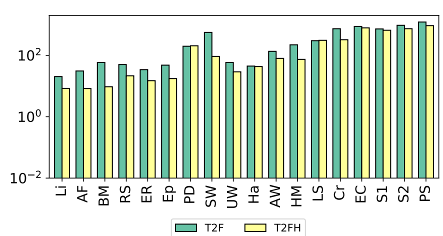
Figure 1: Scalability, varying V, S and N



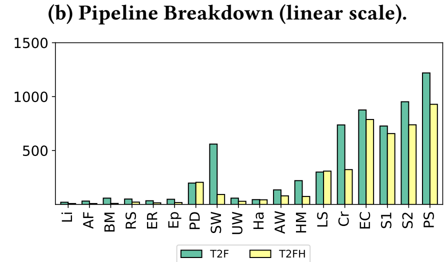
(a) Pipeline Breakdown (log scale).



(b) Pipeline Breakdown (linear scale).



(c) Custom parallelization (log scale).



(d) Custom parallelization (linear scale).

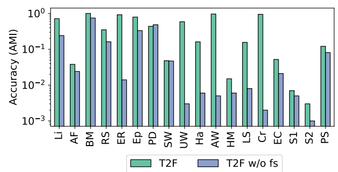
Figure 2: Efficiency analysis.

1.4.3 Importance of the features in the clustering task. This experiment evaluates whether a feature-based clustering approach is more effective than an approach based on raw data. To this end, we run Time2Feat in the unsupervised mode by performing the clustering computation with the same techniques used in the previous experiment (Hierarchical, KMeans, and Spectral), and we compare the accuracy obtained (in terms of AMI) with the one obtained by the application of the same clustering technique to the raw datasets.

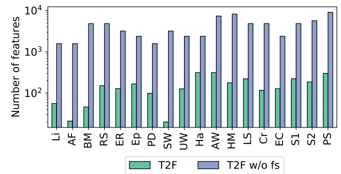
REFERENCES

[1] Anthony J. Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn J. Keogh. 2018. The UEA multivariate time series classification archive, 2018. *CoRR* abs/1811.00075 (2018).

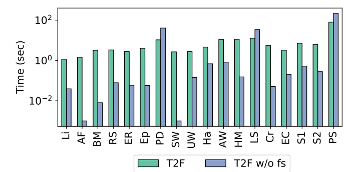
[2] Dino Ienco and Roberto Interdonato. 2020. Deep Multivariate Time Series Embedding Clustering via Attentive-Gated Autoencoder. In *PAKDD 2020 - 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (Advances in Knowledge Discovery and Data Mining)*. Singapore, Singapore. https://hal.inrae.fr/hal-02923636



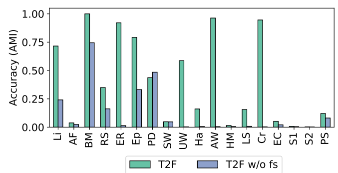
(a) Accuracy (log scale).



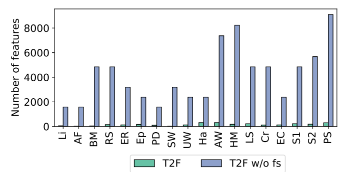
(b) Interpretability (log scale).



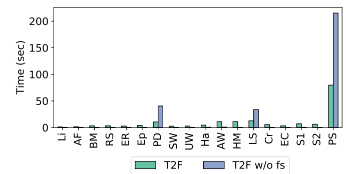
(c) Efficiency (log scale).



(d) Accuracy (linear scale).



(e) Interpretability (linear scale).



(f) Efficiency (linear scale).

Figure 3: Removing the Features Selection from the pipeline.

Table 5: Accuracy (AMI) varying the clustering techniques. In bold, the best result per dataset.

Dataset	$T2F_0$			$T2F_2$			$T2F_5$		
	Hier.	KMeans	Spec.	Hier.	KMeans	Spec.	Hier.	KMeans	Spec.
Li	0.716	0.711	0.627	0.728	0.691	0.709	**0.73**	0.722	0.705
AF	0.038	0.007	-0.001	0.028	0.047	0.047	**0.238**	0.192	0.188
BM	**1**	**1**	0.992	0.977	0.961	0.902	**1**	**1**	0.993
RS	0.35	0.359	0.371	0.559	0.578	0.612	**0.71**	0.649	0.663
ER	0.921	**0.955**	0.925	0.801	0.819	0.79	0.826	0.824	0.804
Ep	0.792	0.874	0.528	**0.896**	0.839	0.8	0.882	0.867	0.793
PD	0.437	0.639	0.377	0.752	0.727	0.651	**0.784**	0.715	0.688
SW	0.048	0.071	-0.004	0.038	0.074	0.167	0.231	**0.327**	0.256
UW	0.587	0.566	0.454	0.555	0.541	0.511	**0.59**	0.539	0.537
HM	0.161	0.153	0.014	0.325	0.302	0.277	**0.349**	0.325	0.289
AW	**0.963**	0.945	0.807	0.921	0.918	0.89	0.927	0.903	0.899
HM	0.015	0.011	0.005	0.021	0.048	0.037	0.069	**0.089**	0.062
LS	0.156	0.146	0.041	0.293	0.315	0.037	**0.333**	0.332	0.051
Cr	0.946	0.907	0.787	**0.984**	0.956	0.95	0.974	0.96	0.967
Ec	0.052	0.056	0.049	0.065	0.066	0.06	**0.121**	0.094	0.106
S1	0.007	0.019	0.002	0.397	**0.419**	0.387	0.382	0.391	0.379
S2	0.003	0	0	0.008	0.015	0.021	0.015	0.024	**0.035**
PS	0.121	0.143	0.143	0.2	**0.211**	**0.211**	0.201	0.208	0.208

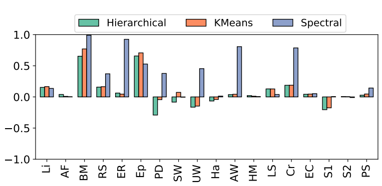


Figure 4: Difference (AMI) between feature-based and raw data clustering.

[3] Yanfei Kang, Rob J. Hyndman, and Feng Li. 2020. GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. *Stat. Anal. Data Min.* 13, 4 (2020), 354–376.

[4] Hailin Li. 2019. Multivariate time series clustering based on common principal component analysis. *Neurocomputing* 349 (2019), 239–247.

[5] Hailin Li, Chunpei Lin, Xiaoji Wan, and Zhengxin Li. 2019. Feature representation and similarity measure based on covariance sequence for multivariate time series. *IEEE Access* 7 (2019), 67018–67026.

[6] Laura Manduchi, Matthias Hüser, Gunnar Rätsch, and Vincent Fortuin. 2019. Variational PSOM: Deep Probabilistic Clustering with Self-Organizing Maps. *CoRR* abs/1910.01590 (2019).

[7] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[8] Xuan Vinh Nguyen, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary?. In *ICML (ACM International Conference Proceeding Series)*, Vol. 382. ACM, 1073–1080.

[9] Miguel A. Atencia Ruiz, Claudio Gallicchio, Gonzalo Joya, and Alessio Micheli. 2020. Time Series Clustering with Deep Reservoir Computing. In *ICANN (2) (Lecture Notes in Computer Science)*, Vol. 12397. Springer, 482–493.

[10] Chenxiao Xu, Hao Huang, and Shinjae Yoo. 2021. A Deep Neural Network for Multivariate Time Series Clustering with Result Interpretation. In *IJCNN*. IEEE, 1–8.