Unsupervised Feature Recommendation using Representation Learning

Anish Datta
TCS Research
TATA Consultancy Services
Kolkata, India
anish.datta@tcs.com

Soma Bandyopadhyay
TCS Research
TATA Consultancy Services
Kolkata, India
soma.bandyopadhyay@tcs.com

Shruti Sachan
TCS Research
TATA Consultancy Services
Kolkata, India
sachan.shruti@tcs.com

Arpan Pal
TCS Research
TATA Consultancy Services
Kolkata, India
arpan.pal@tcs.com

***Abstract*—Today's world extensively depends on analytics of high dimensional sensor time-series, and, extracting informative representation. Sensor time-series across various applications such as healthcare and human wellness, machine maintenance etc., are generally unlabelled, and, getting the annotations is costly and time-consuming. Here, we propose an unsupervised feature selection method exploiting representation learning with a choice of best clustering and recommended distance measure. Proposed method reduces the feature space, to a compressed latent representation, known as Auto-encoded Compact Sequence of features, by retaining the most informative parts. It further selects a set of discriminative features, by computing the similarity / dissimilarity among the features in latent space using the recommended best distance measure. We have experimented using diverse time-series from UCR Time Series Classification archive, and observed, proposed method consistently outperforms state-of-the-art feature selection approaches.**

***Index Terms*—feature selection, unsupervised learning, representation learning**

I. Introduction

High dimensional time-series collected from diverse sensors, are being increasingly used for performing analytics in real-world applications. Commonly, non-informative features exist in the extracted feature set from time-series, and hence, causes overfitting. Reduction of feature space with optimal set of features is a prime need in machine learning(ML).

Feature recommendation is the process of selecting a smaller set of extracted features, which, when combined captures most informative parts of the dataset. Annotations are required in many established scenarios, to select appropriate features, with maximum class separation. However in real world, annotations are scarce, labelling by experts is costly and manual labelling is tedious and time-consuming. Thus, unsupervised feature derivation is an important need across diverse applications like cardiac abnormality detection (healthcare) [1], fault detection in machinery analysis etc.

In this work, we propose an unsupervised feature recommendation method exploiting representation learning [2] and the choice of best distance measure and clustering. The key contributions of proposed method are:

(1) **Learning latent representation of features:** Proposed method computes the discriminative properties of the features in an highly informative representation space. A low-dimensional representation (Auto-Encoded Compact Sequence - AECS) [3] of the features are learned using a multi-layer auto-encoder which captures its most important parts.

(2) **Consistent clustering with recommended distance measure in latent space:** Subsequently, we perform agglomerative hierarchical clustering on the learned representations to form consistent clusters i.e. iteratively forming clusters until it can't be broken into further subclusters. A choice of distance measure is recommended among Chebyshev, Manhattan and Mahalanobis distance for forming the clustering, ensuring generalization across diverse datasets.

(3) **Selection of most important features:** Corresponding to each consistent cluster, redundant features are removed using pairwise correlation score, and an triangular distance matrix is computed on the remaining feature representations, exploiting the recommended distance measure, based on which discriminative features are chosen.

(4) **Extensive experimentation on diverse time-series:** We have experimented using 15 time-series from UCR Time Series Classification Archive [4] corresponding to different application domains. We observed, proposed method outperforms both unsupervised and supervised state-of-the-art (SoA) feature selection approaches.

II. Related Works

Feature selection has been an important area of research due to its impact in diverse application domains. Broadly, feature selection approaches can be classified into two types - (1) Wrapper based and (2) Filter based methods. Wrapper techniques like Forward feature selection [5] and Backward feature elimination [6] employs a specific machine learning algorithm to select a subset of features which maximizes the performance of the algorithm. On the other hand, filter based methods such as mRMR [7], focuses on each features independently to remove or filter out the features having low discriminative properties or are redundant w.r.t other features. Conditional Mutual Information Maximization criterion (CMIM) [8] is another such method, which selects features based on maximization of conditional mutual information along with the class information. Correlation based feature selection (CFS) [9] selects features having maximum correlation with respect to the classes to predict, and minimum intercorrelation among the other features. Another method focusses on selection of

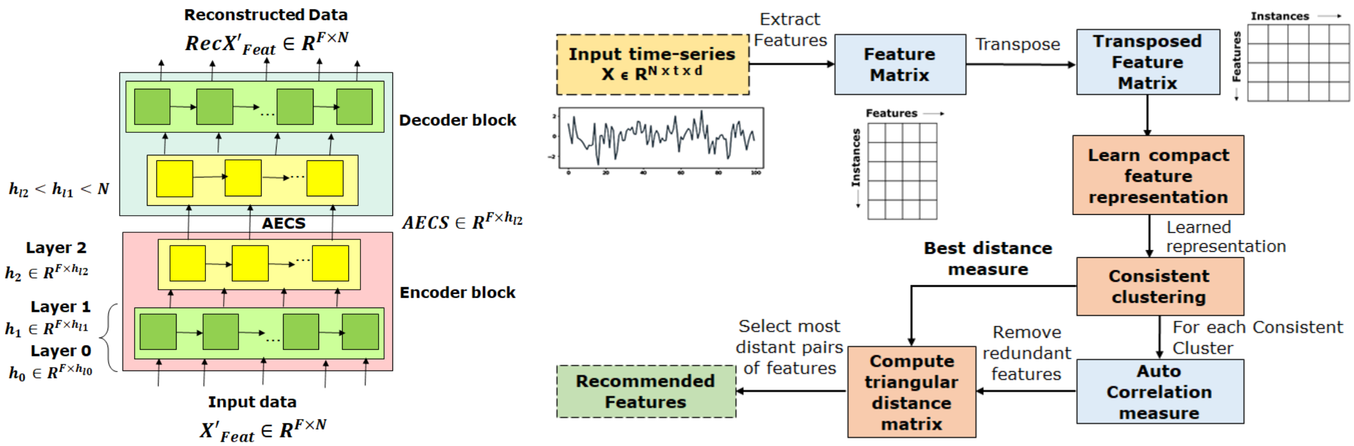


Fig. 1: (a) Schematic architecture of multi-layer Seq-2-Seq auto-encoder for computing feature representation (left) and (b) Framework for proposed unsupervised feature selection method (right)

a subset of features constrained to maximization of the Joint mutual information (JMI) [10] between the selected features.

In recent times, there has been growing research in unsupervised feature selection techniques as well. Multi-Cluster Feature Selection (MCFS) [11], an unsupervised feature selection method, computes the correlation between different features exploiting spectral analysis, which becomes essential in preserving the multi-cluster structure of the data. Spectral Feature Selection (SFS) [12] constructs a pairwise similarity graph and selects the features from the spectrum of the graph. Another method, Nonnegative Discriminative Feature Selection (NDFS) [13] performs spectral clustering and feature selection, in conjunction, to select the optimal subset of features without the use of any labels. Subspace clustering guided unsupervised feature selection (SCUFS) [14], is another method, which learns multi-subspace structure of data in a global similarity matrix, using representation based subspace learning.

III. Methodology

Proposed feature selection method can be used for any sets of features extracted from the time-series, and hence, can be generalized across multiple domains.

As an exemplary feature set, we experiment using 392 rich features extracted from the raw time-series, known as Signal Property based Generic Features (SPGF) [15]. Diverse Temporal, Spectral and Wavelet based features are extracted from the raw time-series which encapsulates its morphology, statistics, randomness and regularity. The features extracted are stored in form of a matrix $X_{Feat} \in \mathcal{R}^{N \times F}$, where N indicates total instances and F indicates the number of features. In this section, we describe our unsupervised feature recommendation method in details.

A. Consistent clustering with recommended distance measure in latent space

1) Learning latent representation of features: Feature matrix X_{Feat} is transposed, such that, the features are represented as different instances in the dataset i.e $X'_{Feat} \in \mathcal{R}^{F \times N}$. Using X'_{Feat}, an auto-encoded compact sequence (AECS) [3] [16] for each of the features are learned using a Seq-2-Seq LSTM multi-layer auto-encoder [17] [18] as shown in Fig. 1(a). Assuming the length of representation to be l, a latent feature matrix $AECS_{Feat} \in \mathcal{R}^{F \times l}$ is formed, where each row represents a highly informative representation of each feature.

2) Forming consistent feature clusters: After the compact representation for each feature is learned, agglomerative hierarchical clustering [19] is performed on it using a method to find the best choice of distance measure. Three different distance measures - Chebyshev, Manhattan [20] and Mahalanobis [21] have been used to evaluate the distance between any two time-series for hierarchical clustering. To find the best distance measure among the above three, and its corresponding clustering, we use an internal clustering measure, Modified Hubert Statistic ($\mathcal{T}$) [22].

For any particular dataset, extracting the knowledge of the number of clusters to be formed, is a very challenging task. To address this, we devise a mechanism called consistent clustering to discover the optimal number of groups of features without any annotations or prior knowledge. We build upon the assumption, that the clusters which are inherently present in a dataset, cannot be broken further into subgroups. It iteratively groups the features using the above mentioned clustering technique, incrementing the number of groups to be formed by 1 in each iteration. If at any iteration, we observe the majority of the features are retained in the same group as in the previous iteration and minimal number of features have been separated to form a new group, then it signifies that the groups can not be divided further optimally and the process is stopped. We define a threshold T as the stopping criterion, to track if the new cluster formed at an iteration has lower than T fraction of the instances of the dataset. It infers the new group formed is very small in size, and the set of previous groups are returned as the consistent clusters.

B. Selection of most important features

1) Removing redundant features using correlation coefficient: In each of the consistent groups of features formed, we first compute the Pearson's correlation coefficient [23] on the learned representation AECS for each pair of features. Considering x_{AECS} and y_{AECS} as learned representation of features x and y grouped in the same cluster, Pearson's correlation coefficient is computed as:

$$P(x,y) = \frac{\sum(x_{AECS} - m_x)(y_{AECS} - m_y)}{\sqrt[2]{\sum(x_{AECS} - m_x)^2 \sum(y_{AECS} - m_y)^2}} \quad (1)$$

where m_x and m_y are the mean of x_{AECS} and y_{AECS} respectively.

If a pair of features have a correlation coefficient above a predefined threshold $Corr_{thresh}$, we infer the features are similar to each other, and, discard one of the features chosen randomly from the pair. This helps to reduce the redundancy in each of the feature groups formed.

2) Selection of discriminative features based on choice of best distance measure: After discarding the redundant features using correlation coefficient, from each of the groups, we exploit the best distance measure chosen during HC-AECS to select the list of recommended features. A triangular distance matrix is computed by finding the distance between each pair of AECS of the remaining features (after removing redundant features) for each cluster separately. Suppose $[{f^c}_1, {f^c}_2, {f^c}_3, ..., {f^c}_n]$ be the remaining features in consistent cluster c. The triangular distance matrix $Dist\_mat_c$ can be computed as

$$Dist\_mat_c[i,j] = d_{best}(AECS({f^c}_i), AECS({f^c}_j)), \quad (2)$$

where d_{best} is the best distance measure recommended by HC-AECS and $AECS(x)$ denotes the learned representation of the corresponding feature x.

In the triangular distance matrix formed for each of the groups, pairs of features having highest distance between them i.e. most distant are selected in the list of recommended features. This ensures maximum separation between selected features, thus minimizing redundancy between them. The number of such distant pairs (n) to be chosen from each group is taken as a hyperparameter. For example, if n is 3 it suggests the top 3 distant pairs of features from each group are added in the list of recommended features. The detailed algorithm and complete framework for proposed method are depicted in Algorithm 1 and Fig. 1(b) respectively.

IV. Experimental Analysis

A. Dataset Description

We evaluate our proposed method on 15 uni-variate time-series from UCR Time Series Classification Archive [4], spread across diverse application domains collected from sensors like ECG, Camera, Process control sensors etc. The number of timesteps of the time-series datasets considered varies widely from 80 to 500. Furthermore, we experiment with both binary and multi-class time-series, where number of classes ranges from 2 to 10.

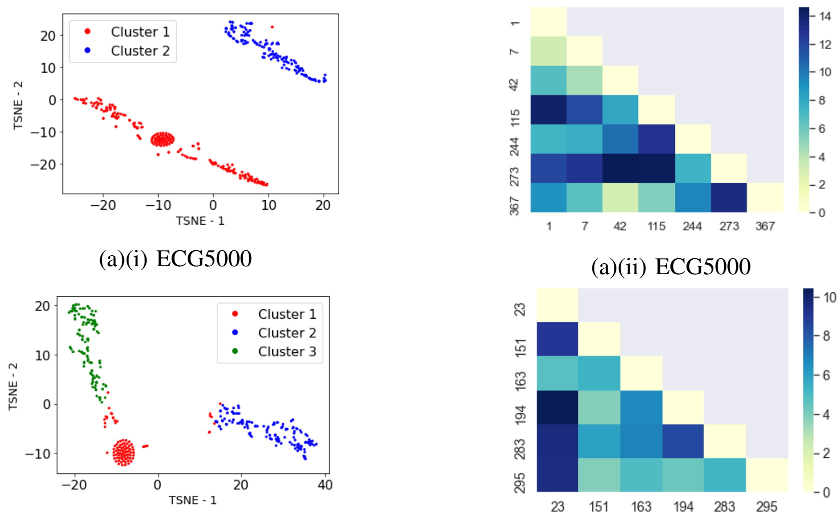
Algorithm 1: Unsupervised Feature Selection

Input : X$\in \mathbb{R}^{N \times t \times d}$: Input time-series,
$Corr_{thresh}$: Threshold for correlation score
n: Number of pairs chosen from each cluster
Output: $Feat_{Rec}$: List of recommended features
Function ***Feature_Selection*** *X*
 ▷ Compute complete feature set for X
 $X_{Feat} \leftarrow$ extract_features(X);
 ▷ Transpose feature matrix
 $X'_{Feat} \leftarrow$ Transpose(X_{Feat});
 ▷ Learn compact feature representation (AECS)
 $AECS_{Feat} \leftarrow$ Multilayer auto-encoder (X'_{Feat});
 ▷ Consistent clustering on $AECS_{Feat}$ to group features based on their similarity in latent space
 $\{C_1, ., C_k\}, d_{best} \leftarrow$ Consistent_Cls($AECS_{Feat}$)
 ▷ Initialize list of recommended features
 $Feat_{Rec} \leftarrow \{\}$;
 forall *Feature cluster* C_i **do**
  ▷ Remove features in C_i having correlation $\geq$ $Corr_{thresh}$ w.r.t other features.
  ▷ Compute distance matrix having distance between each pair of remaining features in C_i using best distance measure d_{best}
  $Dist\_mat \leftarrow$ distance_matrix(C_i, d_{best});
  ▷ Add n most distant feature pairs from $Dist\_mat$ in $Feat_{Rec}$
 end
 return $Feat_{Rec}$;
end

B. Unsupervised Feature Selection

1) Formation of consistent feature clusters:: After extraction of features from the raw time-series, the feature matrix is transposed and a compact representation of all the 392 features is computed using a multi-layer Seq-2-Seq auto-encoder. For experimentation, the length of the AECS representation (l) is considered 12. We use Adam optimizer [24] with a learning rate of 0.004, and the model is run on 100 epochs. Consistent clustering is performed on the resulting latent representation (AECS) matrix ($AECS_{Feat} \in R^{392 \times 12}$), thus producing a number of feature clusters. Fig. 2(a)(i) and (b)(i) depicts the T-SNE plots [25] showing the consistent clusters of features formed in latent space for datasets ECG5000 and MiddlePhalanxTW respectively.

2) Selection of features:: For removing redundant features using correlation coefficient from each of the consistent clusters, we chose a threshold 0.96 i.e, if a pair of feature representations have correlation score greater than 0.96, one of features are discarded randomly. Subsequently, in each of the consistent clusters, a triangular distance matrix is computed for the remaining features using the best distance measure chosen by HC-AECS. We have considered the number of most distant pairs(n) from the distance matrix of each consistent cluster to select the discriminating features. In our experimentation, we



(a)(i) ECG5000 (a)(ii) ECG5000

(b)(i) MiddlePhalanxTW (b)(ii) MiddlePhalanxTW

Fig. 2: (a) (i) TSNE plot of AECS of 392 SPGF features and (ii) Heat map showing the triangular distance matrix for consistent cluster 1 of ECG5000 and (b) (i) TSNE plot of AECS of 392 SPGF features and (ii) Heat map showing the triangular distance matrix for consistent cluster 1 of MiddlePhalanxTW

TABLE I: Performance comparison (in terms of accuracy) of recommended features from proposed method and different SoAs

Sensor	Dataset	Unsupervised methods					Supervised methods				Rank
		Proposed	**MCFS**	**SFS**	**NDFS**	**SCUFS**	**CMIM**	**mRMR**	**CFS**	**JMI**	**(Proposed)**
Camera (Imagine outline as time-series)	ProximalPhalanxOAG	0.795	0.766	0.488	**0.849**	0.834	0.776	0.82	0.765	0.805	5
	ProximalPhalanxOC	0.8	**0.835**	0.684	0.821	0.8	0.819	0.787	0.684	0.807	5
	MedicalImages	**0.699**	0.650	0.514	0.637	0.629	0.661	0.572	0.579	0.546	1
	DistalPhalanxOAG	0.778	0.803	0.643	0.733	0.805	0.755	0.77	0.753	**0.813**	4
	MiddlePhalanxTW	0.591	0.566	0.401	**0.609**	0.529	0.579	0.586	0.531	0.559	2
	DistalPhalanxOC	0.728	0.635	0.630	0.760	0.766	**0.787**	0.76	0.688	0.765	6
	DistalPhalanxTW	0.680	0.656	0.205	**0.780**	0.715	0.73	0.735	0.7	0.755	7
	ProximalPhalanxTW	0.680	0.735	0.450	0.745	**0.79**	0.713	0.718	0.68	0.73	7
	MiddlePhalanxOC	0.600	0.558	**0.647**	0.580	0.62	0.528	0.547	0.58	0.508	3
	Yoga	0.680	0.638	0.536	0.680	0.629	**0.72**	0.593	0.713	0.567	3
Food Spectrograph	Strawberry	**0.920**	0.868	0.643	0.897	0.871	0.819	0.808	0.786	0.829	1
ECG	ECG5000	0.880	**0.891**	0.584	0.878	0.879	0.887	0.88	0.87	0.879	3
Vibration	FordB	0.855	0.793	0.512	0.772	0.797	0.536	**0.86**	0.831	0.664	2
Simulated	ChlorineConcentration	0.590	**0.616**	0.533	0.612	0.572	0.574	0.586	0.487	0.567	3
	SyntheticControl	0.920	0.647	0.167	0.630	0.633	**0.933**	0.7	0.903	0.85	2
Average Accuracy		**0.746**	0.710	0.509	0.732	0.725	0.721	0.715	0.703	0.710	Average Rank = 3.6
Reduction in avg. accuracy than prop. method		-	0.036	0.237	0.014	0.022	0.025	0.031	0.043	0.036	
Average Rank		**3.6**	4.4	8.3	3.7	4.2	4.2	4.7	6.2	5.1	

have considered $n = 3$. If the value of n is increased, higher number of features are recommended by proposed method, and hence can be used as a hyper-parameter to control the number of important features to be selected. We have experimented using the 392 SPGF features extracted from the raw time-series as described above. Examples of lower triangular distance matrix in form of heatmap are shown in Fig. 2(a)(ii) and (b)(ii) for datasets ECG5000 and MiddlePhalanxTW corresponding to a consistent cluster (cluster 1) across the feature members of cluster 1. For ECG5000, we observe feature pairs (273,367), (115,273) and (1,115) are most distant, and hence, $\{1, 115, 273, 367\}$ are returned as the most representative features from cluster 1. Here feature 1 indicates *Mean on Approximation coefficients of First Level Discrete Wavelet Transform (DWT)* [26], 115 indicates *Zero Crossing Rate* [27] *of time domain signal*, 273 indicates *Standard deviation of windowed Zero Crossing Rate of DWT (First level approximation coefficients of DWT)* and 367 indicates *Mean of windowed Skewness in frequency domain*.

C. Results

For evaluating the discriminative properties of the selected features, we consider a classification setting where only the recommended features are used for training and inference. The training set is used for feature selection and training a TreeBagger classifier [28] using the selected features, while the testing set is reserved for inferencing only. We compare proposed method with four state-of-the-art unsupervised feature selection methods - MCFS, SFS, NDFS and SCUFS. For fair comparison, if k be the number of features recommended by proposed method, we consider top-k ranked features by the SoA feature selection techniques, as they do not recommend the number of features to be selected. The inference results (in accuracy) on the test set for proposed method and the SoAs are provided in Table I. We observe proposed method outperforms all the benchmark approaches both in terms of average accuracy within a range of 23.7 % to 1.4 %, and also average rank.

Furthermore, we also compare our approach with supervised selection techniques like mRMR, CMIM, CFS and JMI. We observe proposed method outperforms even supervised feature selection techniques, which uses label information, within a range of 4.3 % to 2.5 %, across the 15 datasets. Further, we observe proposed method achieves the lowest average rank, among all other methods, where rank 1 denotes highest performance. In another comparison, w.r.t a supervised setup considering entire feature set of SPGF with Support Vector Machine (SVM) Grid search [29] as classifier and recommended features of our method using same classifier and identical hyper parameters with the same datasets we noticed an improvement in 8 out of 15 datasets with 2.1% increase in average accuracy.

V. Conclusion

Here, we have presented a robust unsupervised feature selection method using representation learning and forming hierarchical clustering with a choice of best distance measure chosen among Chebyshev, Manhattan and Mahalanobis distance. The distance measure recommended by best clustering selects the important features considering their relevance in terms of their respective pairwise distance inside the consistent clusters formed using the compact latent representation. We have performed extensive analysis considering real world time-series from different application domains like health-care, machine maintenance etc. We have also illustrated our method using signal processing and morphology aware feature set. Experimental results depict selected features obtained from proposed method outperforms benchmark unsupervised feature selection methods, as well as, supervised feature selection techniques.

References

[1] S. Bandyopadhyay, A. Ukil, C. Puri, R. Singh, A. Pal, K. Mandana, and C. Murthy, "An unsupervised learning for robust cardiac feature derivation from ppg signals," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 740–743.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] S. Bandyopadhyay, A. Datta, and A. Pal, "Hierarchical clustering using auto-encoded compact representation for time-series analysis," in *Learning Data Representation for Clustering Workshop, IJCAI-PRICAI*, 2020.

[4] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The ucr time series archive," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.

[5] D. Ververidis and C. Kotropoulos, "Sequential forward feature selection with low computational cost," in *2005 13th European Signal Processing Conference*. IEEE, 2005, pp. 1–4.

[6] D. Kostrzewa and R. Brzeski, "The data dimensionality reduction in the classification process through greedy backward feature elimination," in *International Conference on Man–Machine Interactions*. Springer, 2017, pp. 397–407.

[7] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The journal of machine learning research*, vol. 13, pp. 27–66, 2012.

[8] F. Fleuret, "Fast binary feature selection with conditional mutual information." *Journal of Machine learning research*, vol. 5, no. 9, 2004.

[9] M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper." in *FLAIRS conference*, vol. 1999, 1999, pp. 235–239.

[10] H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," *Advances in neural information processing systems*, vol. 12, 1999.

[11] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 333–342.

[12] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 1151–1157.

[13] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 26, no. 1, 2012.

[14] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognition*, vol. 66, pp. 364–374, 2017.

[15] A. Ukil, P. Malhotra, S. Bandyopadhyay, T. Bose, I. Sahu, A. Mukherjee, L. Vig, A. Pal, and G. Shroff, "Fusing features based on signal properties and timenet for time series classification." in *ESANN*, 2019.

[16] S. Bandyopadhyay, A. Datta, and A. Pal, "Automated label generation for time series classification with representation learning: Reduction of label cost for training," in *Weakly Supervised Representation Learning Workshop, IJCAI*, 2021.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[20] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 275–309, 2013.

[21] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, vol. 50, no. 1, pp. 1–18, 2000.

[22] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[23] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[26] M. J. Shensa *et al.*, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[27] F. Gouyon, F. Pachet, O. Delerue *et al.*, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*, vol. 5. Citeseer, 2000, p. 16.

[28] C. D. Sutton, "Classification and regression trees, bagging, and boosting," *Handbook of statistics*, vol. 24, pp. 303–329, 2005.

[29] P. Lameski, E. Zdravevski, R. Mingov, and A. Kulakov, "Svm parameter tuning with grid search and its impact on reduction of model overfitting," in *Rough sets, fuzzy sets, data mining, and granular computing*. Springer, 2015, pp. 464–474.