Classification of fMRI data using Dynamic Time Warping based functional connectivity analysis

Regina Meszlényi[1,3], Ladislav Peska[2], Viktor Gál[3], Zoltán Vidnyánszky[3], Krisztian Buza[3]

[1]Department of Cognitive Science, Budapest University of Technology and Economics, Budapest, Hungary

[2]Faculty of Mathematics and Physics Charles University in Prague Czech Republic

[3]Brain Imaging Centre Research Centre for Natural Sciences Hungarian Academy of Sciences Budapest, Hungary

***Abstract*—The synchronized spontaneous low frequency fluctuations of the BOLD signal, as captured by functional MRI measurements, is known to represent the functional connections of different brain areas. The aforementioned MRI measurements result in high-dimensional time series, the dimensions of which correspond to the activity of different brain regions. Recently we have shown that Dynamic Time Warping (DTW) distance can be used as a similarity measure between BOLD signals of brain regions as an alternative of the traditionally used correlation coefficient. We have characterized the new metric's stability in multiple measurements, and between subjects in homogenous groups. In this paper we investigated the DTW metric's sensitivity and demonstrated that DTW-based models outperform correlation-based models in resting-state fMRI data classification tasks. Additionally, we show that functional connectivity networks resulting from DTW-based models as compared to the correlation-based models are more stable and sensitive to differences between healthy subjects and patient groups.**

Keywords*—*fMRI; functional connectivity networks; dynamic time warping; classification

I. Introduction

The functional organization of the human brain has long been studied with task-based functional magnetic resonance imaging (fMRI) measurements, however it has been shown that the synchronised spontaneous low frequency fluctuations of the BOLD signal during rest also represent the functional networks of brain areas [1]–[3]. Traditionally resting-state brain networks are analysed with techniques that imply static zero-lag linear dependence between brain regions, e.g. as the fMRI measurement results in a high-dimensional time series (one 1D time-series for each volume pixels a.k.a. voxels of the brain) the strength of functional connectivity between any pair of voxels is usually characterized with the Pearson correlation coefficient of the two measured signal [1]. Other methods for revealing functional networks like independent component analysis [4] are similarly popular in the neuroimaging community, yet they still rely on measures of linear dependence.

On the other hand, growing number of neuroimaging studies suggest that functional networks display dynamic changes in connectivity strength [5]–[7], as well as varying phase difference (nonzero time-lag) between regions [8]. To address these issues we proposed to use Dynamic Time Warping (DTW) distance [9] as an alternative measure of similarity between BOLD signals [10]. We were able to show that DTW results in more stable functional connectivity than correlation, in multiple measurements and with different preprocessing strategies, since DTW can effectively handle non-stationary processes.

Besides the fact that resting-state functional connectivity provides insight into the functional organization of the human brain, it also has great potential as a biomarker of several mental disorders. It has been shown that not only somewhat trivial differences like age-groups and gender can be classified via functional connectivity strength [11], [12], but there are encouraging results in case of mental disorders like Alzheimer's disease or ADHD (Attention Deficit Hyperactivity Disorder) as well [13]–[15].

Based on the results of [10] we hypothesised that DTW based resting-state functional connectivity can be an applicable input of classification algorithms. In this paper we compare classification performances based on resting-state functional connectivity measured with DTW and correlation, to further validate our claim that the Dynamic Time Warping distance is indeed a suitable descriptor of connectivity between brain regions. For the comparison with other connectivity measures such as cross-correlation, see [10].

As models used to classify the fMRI data can be interpreted in terms of brain networks, we compare the resulting brain networks and demonstrate that DTW-based networks may reveal differences between the functional connectivity patterns of healthy subjects and ADHD patients. We discuss networks based on DTW and correlation, and illustrate that DTW-based networks may be preferable to correlation-based networks.

The rest of the paper is organized as follows. Section II summarize the methodical background relevant to understand the paper, Section III presents our results followed by the discussion of the results in Section IV.

II. MATERIALS AND METHODS

A. Data and preprocessing

In order to assist reproducibility and to perform classification of fMRI data according to a standard protocol, we downloaded a preprocessed public resting-state fMRI database from the 1000 Functional Connectomes Project, Addiction Connectome Preprocessed Initiative. In our study we used the MTA 1 dataset with the ANTS registered, no scrubbing, no global signal regression preprocessing pipeline. Detailed description of the preprocessing strategy is available at the homepage of the dataset [16]. The downloaded dataset contains 126 subjects' resting-state data as well as phenotypic information including gender (25 females, 101 males), and childhood diagnosis for ADHD (40 subjects with positive, 86 with negative diagnosis). For a connectivity based classification we used an atlas of 90 functional regions of interest (ROI) [17] to obtain 90 functionally meaningful averaged BOLD signals per subject. From this 90 time series we calculated full connectivity matrices with Pearson correlation as well as with DTW. Possibly due to the registration process used in the published preprocessing, one ROI (Basal Ganglia 4) included no meaningful measurement data for any of the 126 subjects, therefore we used the remaining 89 ROIs, resulting in 3912 individual connectivity features for classification.

B. Dynamic Time Warping

Dynamic Time Warping (DTW) is a distance measure between time series that takes potential shifting and elongations into account when comparing two time series. Although, it was originally designed for speech recognition [9], in the last decades, DTW was shown to work surprisingly well for time series classification [18], [19], thus it became one of the most prominent time series distance measures in the machine learning community, see e.g. [20] and the references therein.
DTW is an edit distance, i.e., when comparing two time series, it calculates the "cost" of transforming one of the time series into the other one.
Calculation of the DTW distance of two time series of length l_1 and l_2, can be implemented as filling-in the entries of an l_1 x l_2 matrix. This is illustrated in Fig. 1.We consider the first time-series to be written on the left side of the matrix, whereas the second time series is considered to be written on the top of the matrix. Each entry of the matrix corresponds to the distance between two prefixes of the time-series. The entries can be filled-in column-by-column and row-by-row, beginning with the first row in the first column, followed by the second, third, etc. rows in the first column. Once a column is filled, we begin with filling-in the next column. In order to fill an entry $D(i,j)$ of the matrix, we use the following rule:

$$D(i,j) = \|t_1(i)\text{-}t_2(j)\| + \min\big(D(i\text{-}1,j\text{-}1), D(i\text{-}1,j), D(i,j\text{-}1)\big) \quad (1)$$

where $t_1(i)$ denotes the i-th value in time series t_1 and $t_2(j)$ denotes the j-th value in time series t_2. Once the matrix is filled, the value in the entry $D(l_1,l_2)$ is the DTW-distance of the two time series.

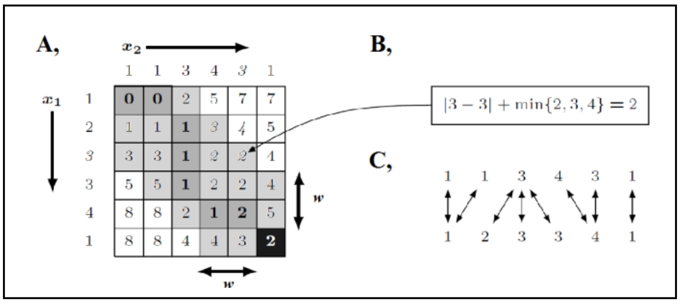

Fig. 1. A, Calculation of DTW distance by filling out the DTW matrix. Elements of x_1 correspond to rows, while elements of x_2 correspond to columns of the matrix; w denotes the size of the warping window, the maximal allowed time-lag between two matched time series element. The optimal warping path is highlighted with dark grey. B, Formula to calculate entry (i,j): distance of $x_1(i)$ and $x_2(j)$ plus the minimum of the matrix entries $(i-1,j)$, $(i-1,j-1)$, $(i,j-1)$. C, Optimal matching of the elements of x_1 and x_2 revealed by the DTW matrix. Adapted with permission from [19].

Once the matrix is filled, starting from $D(l_1,l_2)$, by considering which of the neighboring cells has led to the minimum in Eq. 1., we can construct the warping path, or, equivalently, the matching between the positions of the time series, see Fig. 1C for an example.
In order to speed-up DTW-calculations, it is enough to calculate the cells close to the main diagonal of the matrix [9]. This corresponds to limiting the shifting that is allowed between matched positions of the two time series. In other words: we apply a warping window. For example, when calculating only the marked entries in Fig. 1A, the size of the warping window is $w = 2$.

C. Classifiers

- SVM

The classification based on support vector machines has gained prominent reputation for various machine learning tasks. SVM objective is to find optimal hyperplane $x^T\beta + \beta_0 = 0$, separating training examples of the both classes:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{N} \xi_i \quad (2)$$

subject to $\forall i{:}\ \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$, where N is the number of examples, $x_i \in \mathbb{R}^d$ is the vector of the subject's features, d is the number of features, $y_i \in \{-1,1\}$ is the subject's class, ξ_i is so called slack variable indicating the proportion by which is the train example x_i misclassified and $\lambda \in \mathbb{R}$ is a hyper parameter controlling the tradeoff between hyperplane margin size and misclassification errors.

The described linear SVM can be further refined by enlarging feature space based on some kernel function and thus learning a non-linear separator in the original feature space.

- LASSO regression

The least absolute shrinkage and selection operator is a regularized regression analysis method that performs feature selection, which makes it particularly useful in case of high dimensional datasets [21]. The LASSO's objective is to find the parameter vector $\vec{\theta}$ that minimizes the sum of squared errors and the regularization term:

$$\vec{\theta} = arg\min_{\vec{\theta}} \frac{1}{N} \left\| \vec{y} - \boldsymbol{X}\vec{\theta} \right\|_2^2 + \lambda \left\| \vec{\theta} \right\|_1 \quad (3)$$

where N is the number of examples, $\boldsymbol{X} \in \mathbb{R}^{Nxd}$ matrix contains the cases, d is the number of features, $\vec{y} \in \mathbb{R}^N$ contains the desired output values, $\vec{\theta} \in \mathbb{R}^d$ is the parameter vector, and $\lambda \in \mathbb{R}$ is a hyper parameter controlling the regularization.

LASSO can be considered as a convex relaxation of the best subset selection regression problem, where the regularization term is $\left\| \vec{\theta} \right\|_0$, the number of nonzero entries of the parameter vector. As L_0 is not a norm, since no L_p norm holds the triangle inequality for p<1, the L_1 regularization term used in Lasso regression is the best convex approximation of the subset selection problem.

III. RESULTS

A. Evaluation protocol

In order to show that DTW is able to capture functional connections between regions of the brain, we used DTW distances as features for classification. In particular, we calculated the DTW-distance between each pairs of time series associated with the 89 ROIs. We set the size of the warping window to 100 s corresponding to 50 time-points, since during preprocessing, time-series are bandpass filtered with 0.01 Hz lower cut-off frequency.

We performed experiments according to the leave-one-out cross-validation protocol with two different classification targets, ADHD and Gender, and with two widely-used classifiers, linear SVM and LASSO that were described in Section II. In case of both of these classifiers, in each round of the cross-validation, the value of the hyper parameter λ was determined using the training data only. In particular, we performed an *internal* cross-validation on the training data in order to select the value of λ that maximizes macro-averaged F-measure, i.e., the evaluation metric we used to assess the quality of the models (see below).

In all the aforementioned cases, we calculated F-measure, i.e., the harmonic mean of precision and recall, for both classes and averaged the F-measures of the two classes. This led to a macro-averaged F-measure which we used to assess the quality of the classifiers.

We compared the performance of classifiers using DTW distances as features with that of classifiers using correlation-based features instead of DTW. In order to test if the differences between the performance of these classifiers are statistically significant, we used the binomial test suggested by Salzberg [22].

B. Results of SVM classification

TABLE I. contains results of linear SVM classification for gender and ADHD targets. As can be seen, no statistically significant differences were observed for the gender classification. In fact, both classifiers output a very similar model favoring the major class. For the ADHD classification, linear SVM based on the DTW dataset significantly outperformed linear SVM based on the correlation coefficient.

TABLE I. RESULTS OF LINEAR SVM CLASSIFICATION

Average F-measure	**Classification targets**	
	Gender	***ADHD***
DTW	0.43	0.58
Correlation coefficient	0.47	0.51
Significance level	0.64	4.4E-02

C. Results of LASSO classification

The results obtained with LASSO classification are summarized in Table II. Both in case of gender and ADHD targets DTW based classification outperforms the correlation based method significantly.

TABLE II. RESULTS OF LASSO CLASSIFICATION

Average F-measure	**Classification targets**	
	Gender	***ADHD***
DTW	0.74	0.60
Correlation coefficient	0.42	0.44
Significance level	4.01E-05	1.66E-02

As LASSO accomplishes feature selection, beside the performance measured by macro-averaged F-measure, it is also an important question how many features are selected, and whether the selected features are stable through the outer cycle of the internal leave-one-out cross validation.

We found that LASSO selects more features if connectivity is calculated with DTW, and these features are also more stable than in correlation based classification. The number of features that were selected at least 100 times out of the 126 cycles of cross validation based on DTW connectivity is 89 in case of gender and 70 in case of ADHD targets, while correlation based connectivity yields 61 stable features in gender classification and only 19 in case of ADHD.

The fact that DTW based classification outperforms correlation demonstrates that DTW captures more relevant information of brain connectivity, since correlation based classification tends to overfit the data when it infers from more features.

IV. Discussion

From the results presented in Table I and Table II, it is clearly visible that where meaningful classification was possible, the DTW based classifiers significantly outperformed the ones using correlation features, both with linear SVM and LASSO classifiers.

Additionally to LASSO and linear SVM, we tried further classifiers as well, such as *k*NN and non-linear SVMs. According to our observations, these classifiers did not improve the accuracy of classification compared with the results presented in Table I and Table II. Most likely, the high number of features, which is known under the term of the curse of dimensionality, may explain the aforementioned observation. In particular, in case of SVMs as the training instances are already linearly separable in the original feature space, more complex kernels do not improve classification performance. On the other hand, kNN is known to be incapable to deal with the abundant volume of features, of which some might be irrelevant. For a more detailed discussion of the curse of dimensionality with special focus on *k*NN classifiers, see [23].

In contrast to the aforementioned classifiers, LASSO is known to be useful in case of high dimensional datasets, even with relatively few training examples, since the enforced sparsity reduces the chance of overfitting when it is assumed that only a limited number of dimensions contain relevant information. This is consistent with our observations, according to which LASSO produced the most accurate classification both in case of gender and ADHD.

Furthermore, LASSO's feature selecting property holds additional value in biomedical applications. As in our case features are derived from connections between regions of the brain, the selected subset of relevant features determine a network of brain regions that differs most between the distinguished classes.

Regarding ADHD, the emerging network that is able to differentiate between healthy subjects and patients affected by the disease is particularly interesting from the diagnostic point of view. We can visualize the selected stable features on graphs (see Fig 2.), where the nodes represent the corresponding ROIs, and the edges are the selected features either calculated with correlation or DTW. As we stated in the results section, the number of stable features based on DTW is much higher than in case of correlation, resulting in a network of 68 nodes (ROIs) and 70 edges with DTW and 33 nodes and only 19 edges in case of correlation. Nodes with more edges (high degree) are particularly interesting, as their connectivity patterns influence the classification results the most. The functional ROIs corresponding to graph nodes are visualized in Fig. 3, where the degree of the given node is color coded.

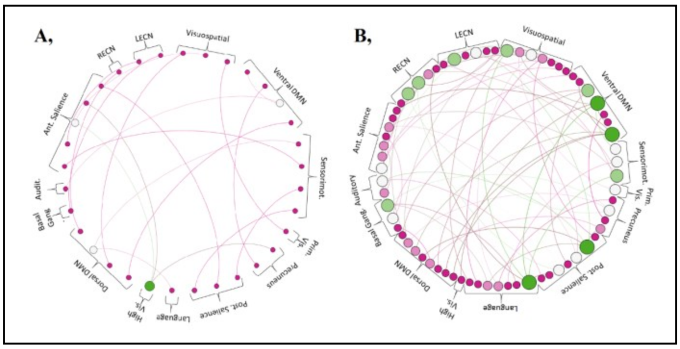

Fig. 2.ADHD networks from LASSO classification A, Graph representation of the ADHD network based on correlation. B, Graph representation of the ADHD network based on DTW distance. The coloring and sizeing of nodes corresponds to the degree of the given node (pink – low degree, green – high degree).

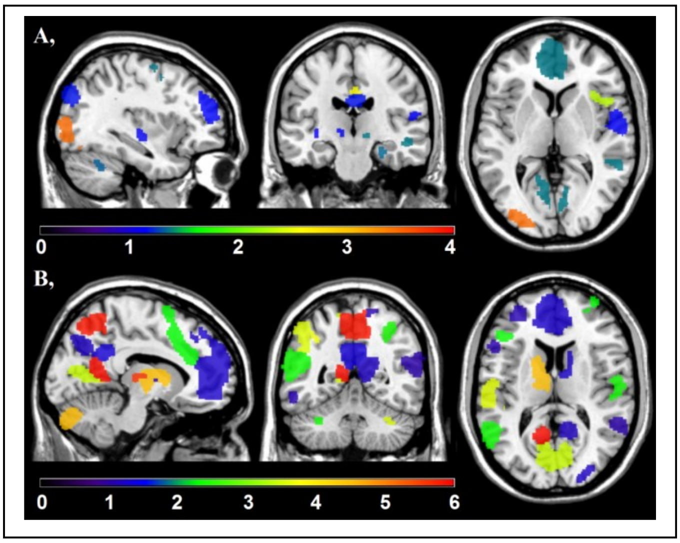

Fig. 3. ADHD networks from LASSO classification A, The nodes of the ADHD network based on correlation mapped back to the functional ROIs. B, The nodes of the ADHD network based on DTW distance mapped back to the functional ROIs. The coloring of the ROIs correspond to the degree of the given node.

Although ROIs found with both correlation and DTW can be explained based on previous ADHD research [24], [25], recent studies emphasized the role of large-scale brain network differences in ADHD [15], [26]. The network emerging from DTW-based features include more regions and more diverse set of connections. This network is able to capture differences between healthy and diseased subjects more efficiently, explaining the higher macro-averaged F-measure reached by the DTW based LASSO classification.

Based on our results we can state that Dynamic Time Warping distance is a suitable measure of functional connectivity strength, since beside its demonstrated stability [10] it also emphasizes group differences resulting in better classification.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Hungarian Brain Research Program (KTIA_13_NAP–A–I/18) to Z. Vidnyánszky. L. Peska was supported by project SVV-2015-260222. K. Buza was supported by the National Research, Development and Innovation Office - NKFIH PD 111710 and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

REFERENCES

[1] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, 'Functional connectivity in the motor cortex of resting human brain using echo-planar MRI', *Magn. Reson. Med.*, vol. 34, no. 4, pp. 537–541, Oct. 1995.

[2] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, A.-M. Dogonowski, M. Ernst, D. Fair, M. Hampson, M. J. Hoptman, J. S. Hyde, V. J. Kiviniemi, R. Kötter, S.-J. Li, C.-P. Lin, M. J. Lowe, C. Mackay, D. J. Madden, K. H. Madsen, D. S. Margulies, H. S. Mayberg, K. McMahon, C. S. Monk, S. H. Mostofsky, B. J. Nagel, J. J. Pekar, S. J. Peltier, S. E. Petersen, V. Riedl, S. A. R. B. Rombouts, B. Rypma, B. L. Schlaggar, S. Schmidt, R. D. Seidler, G. J. Siegle, C. Sorg, G.-J. Teng, J. Veijola, A. Villringer, M. Walter, L. Wang, X.-C. Weng, S. Whitfield-Gabrieli, P. Williamson, C. Windischberger, Y.-F. Zang, H.-Y. Zhang, F. X. Castellanos, and M. P. Milham, 'Toward discovery science of human brain function', *Proc. Natl. Acad. Sci.*, vol. 107, no. 10, pp. 4734–4739, Mar. 2010.

[3] M. D. Fox and M. E. Raichle, 'Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging', *Nat. Rev. Neurosci.*, vol. 8, no. 9, pp. 700–711, Sep. 2007.

[4] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, 'Analysis of fMRI data by blind separation into independent spatial components', *Hum. Brain Mapp.*, vol. 6, no. 3, pp. 160–188, 1998.

[5] D. A. Handwerker, V. Roopchansingh, J. Gonzalez-Castillo, and P. A. Bandettini, 'Periodic changes in fMRI connectivity', *NeuroImage*, vol. 63, no. 3, pp. 1712–1719, Nov. 2012.

[6] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo, D. A. Handwerker, S. Keilholz, V. Kiviniemi, D. A. Leopold, F. de Pasquale, O. Sporns, M. Walter, and C. Chang, 'Dynamic functional connectivity: Promise, issues, and interpretations', *NeuroImage*, vol. 80, pp. 360–378, Oct. 2013.

[7] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, 'Tracking whole-brain connectivity dynamics in the resting state', *Cereb. Cortex N. Y. N 1991*, vol. 24, no. 3, pp. 663–676, Mar. 2014.

[8] C. Chang and G. H. Glover, 'Time-frequency dynamics of resting-state brain connectivity measured with fMRI', *NeuroImage*, vol. 50, no. 1, pp. 81–98, Mar. 2010.

[9] H. Sakoe and S. Chiba, 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Trans. Acoust. Speech Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[10] R. Meszlényi, P. Hermann, K. Buza, V. Gál, and Z. Vidnyánszky, 'Resting state fMRI functional connectivity analysis using Dynamic Time Warping', unpublished, submitted to NeuroImage.

[11] N. U. F. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, K. A. Barnes, J. W. Dubis, E. Feczko, R. S. Coalson, J. R. Pruett, D. M. Barch, S. E. Petersen, and B. L. Schlaggar, 'Prediction of Individual Brain Maturity Using fMRI', *Science*, vol. 329, no. 5997, pp. 1358–1361, Sep. 2010.

[12] L. Tian, J. Wang, C. Yan, and Y. He, 'Hemisphere- and gender-related differences in small-world brain networks: A resting-state functional MRI study', *NeuroImage*, vol. 54, no. 1, pp. 191–202, Jan. 2011.

[13] Y. Liu, K. Wang, C. YU, Y. He, Y. Zhou, M. Liang, L. Wang, and T. Jiang, 'Regional homogeneity, functional connectivity and imaging markers of Alzheimer's disease: A review of resting-state fMRI studies', *Neuropsychologia*, vol. 46, no. 6, pp. 1648–1656, May 2008.

[14] S. Smith, 'Linking cognition to brain connectivity', *Nat. Neurosci.*, vol. 19, no. 1, pp. 7–9, Dec. 2015.

[15] M. D. Rosenberg, E. S. Finn, D. Scheinost, X. Papademetris, X. Shen, R. T. Constable, and M. M. Chun, 'A neuromarker of sustained attention from whole-brain functional connectivity', *Nat. Neurosci.*, vol. 19, no. 1, pp. 165–171, Nov. 2015.

[16] http://fcon_1000.projects.nitrc.org/indi/ACPI/html/acpi_mta_1.html

[17] W. R. Shirer, S. Ryali, E. Rykhlevskaia, V. Menon, and M. D. Greicius, 'Decoding Subject-Driven Cognitive States with Whole-Brain Connectivity Patterns', *Cereb. Cortex*, vol. 22, no. 1, pp. 158–165, Jan. 2012.

[18] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, 'Fast Time Series Classification Using Numerosity Reduction', in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, pp. 1033–1040.

[19] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, 'Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures', *Proc VLDB Endow*, vol. 1, no. 2, pp. 1542–1552, Aug. 2008.

[20] N. Tomasev, K. Buza, K. Marussy, and P. B. Kis, 'Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series', in *Feature selection for data and pattern recognition*, Springer-Verlag, 2015.

[21] R. Tibshirani, 'Regression Shrinkage and Selection via the Lasso', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[22] S. L. Salzberg, 'On Comparing Classifiers: Pitfalls toAvoid and a Recommended Approach', *Data Min Knowl Discov*, vol. 1, no. 3, pp. 317–328, Jan. 1997.

[23] M. Radovanović, A. Nanopoulos, and M. Ivanović, 'Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data', *J Mach Learn Res*, vol. 11, pp. 2487–2531, Dec. 2010.

[24] E. M. Valera, S. V. Faraone, K. E. Murray, and L. J. Seidman, 'Meta-Analysis of Structural Imaging Findings in Attention-Deficit/Hyperactivity Disorder', *Biol. Psychiatry*, vol. 61, no. 12, pp. 1361–1369, Jun. 2007.

[25] S. G. Dickstein, K. Bannon, F. X. Castellanos, and M. P. Milham, 'The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis', *J. Child Psychol. Psychiatry*, vol. 47, no. 10, pp. 1051–1062, Oct. 2006.

[26] F. X. Castellanos and E. Proal, 'Large-Scale Brain Systems in ADHD: Beyond the Prefrontal-Striatal Model', *Trends Cogn. Sci.*, vol. 16, no. 1, pp. 17–26, Jan. 2012.