24th Mediterranean Conference on Control and Automation (MED)
June 21-24, 2016, Athens, Greece

Towards a Fully Automated Tool for Annotation of Phasic Electromyographic Activity

Petros Karvelis, George Georgoulas, Jacqueline A. Fairley, *Senior Member* IEEE, Chrysostomos D. Stylios, *Senior Member* IEEE David B. Rye, and Donald L. Bliwise

Abstract— **Salient muscle activity identification via the phasic electromyographic metric (PEM) in human polysomnograms/sleep studies (PSGs) represent a potential quantitative metric to aid in differentiation between neurodegenerative disorder populations and age-matched controls. A major impairment to the implementation of PEM analysis for clinical assessment of neurodegenerative disorders includes the time consuming aspects for both visual and automated supervised methods, which require exhaustive expert scoring of PEM and non-PEM events. In order to surmount the aforementioned concerns, we propose a semi-supervised classification methodology encased within an easy-to-use graphical user interface (GUI) utilizing an embedded Minimum Description Length (MDL) criterion to automatically classify PEM and non-PEM events based on expert labeling of a single PEM instance. Results indicate that the application of a semi-supervised approach for PEM identification provides an excellent option to reduce the labeling burden within current human PSG muscle activity identification schemes.**

I. Introduction

Identification of salient muscle activity characteristics in human polysomnograms /sleep studies (PSGs) via the phasic electromyographic metric (PEM) provides a candidate quantitative metric to assist in discerning between neurodegenerative disorder populations and age-matched controls [1]. Despite the potential usefulness of PEM detection in assessment of neurodegenerative conditions a standardized processing scheme has yet to be widely accepted in clinical practice [2]. A major barrier to the usage of PEM analysis in clinical practice is attributed to the time consuming aspects for current visual and automated supervised methods, both based on exhaustive expert scoring of PEM and non-PEM events [3]. To decrease the labelling burden of the aforementioned methods, in this work, we propose the implementation of a semi-supervised classification methodology.

In our previous works [3]-[5], the EMG signal was parsed using a one second non-overlapping moving window for the extraction of features capable to characterize segments containing relevant muscle activity. Therefore, candidate PEM events were obtained from all EMG signals. Finally, a supervised classification scheme was implemented to assign candidate PEM events to the PEM or non-PEM class.

Although the aforementioned works documented an enhanced streamlined approach for PEM scoring, compared to visual labelling our current software tool includes a major drawback impairing immediate translation from research application to clinical usage. More specifically, our selection of a supervised method to classify candidate PEM events relied upon the robustness of the selected training data. Thus the performance efficiency of our method depended upon the amount of available labelled data. Furthermore, labelled data is often hard to obtain, expensive, and may frequently require exhaustive human intervention. However, an intermediate exists between Supervised Learning and Unsupervised Learning which is known as Semi-Supervised Learning (SSL) [6]. More importantly, human intervention is significantly reduced in SSL compared to Supervised Learning being that fewer events of interest are required for labelling and more unlabeled instances are accommodated.

Therefore, to build upon our previous findings we extend our work, [7] and [3], based upon the approach proposed in [8] with a subtle modification, described in the next section. However, unlike our previous framework, which required the user to select multiple training events, this revised approach consists of labelling only a single true PEM event that is utilized to detect unlabeled PEM events. Next, we run a stopping criterion using a Minimum Description Length (MDL) ([9], [10]) scheme to classify all unlabeled data as PEM or non-PEM events. Implementation of the MDL within our revised methodology ensures a more robust PEM scoring scheme. Furthermore, the user/expert has the option to accept or disagree/revise any annotations our software proposes. This allows for quick parsing of the signal and immediate correction of possible misclassifications. Finally, the user also has the option to add PEM events, to the final annotation set, that were left undetected by our software.

The rest of the paper is structured as follows: Section II delineates the data collection procedure and processing methods. Section III summarizes our preliminary results, while Section IV concludes the paper along with providing

P. Karvelis is with the Laboratory of Cardiovascular Biology and Biomechanics Laboratory, Cardiovascular Division, University of Nebraska Medical Center, Omaha Nebraska, USA (e-mail: petros.karvelis@unmc.edu).

G. Georgoulas is with the Laboratory of Knowledge and Intelligent Computing, Technological Educational Institute of Epirus, Department of Computer Engineering Arta, Greece, (email: georgoul@gmail.com).

J. A. Fairley, is with the Georgia Tech Research Institute, Sensors & Electromagnetic Applications Laboratory, Atlanta, Georgia, USA (email: jacqueline.fairley@gtri.gatech.edu).

C. D. Stylios, are with the Laboratory of Knowledge and Intelligent Computing, Technological Educational Institute of Epirus, Department of Computer Engineering Arta, Greece (e-mail: stylios@teiep.gr).

David B. Rye and Donald L. Bliwise Emory University, School of Medicine Department of Neurology, Atlanta, Georgia, USA (e-mail: drye@emory.edu, dbliwis@emory.edu).

978-1-4673-8345-5/16/$31.00 ©2016 IEEE

some guidelines for future research and potential improvements.

II. Materials and Methods

A. Data Collection

Data sets utilized in this study were obtained in compliance with Institutional Review Board (IRB) guidelines established by Emory University (Atlanta, Georgia, USA) and sanctioned by the approved IRB00024934 protocol. Polysomnogram (PSG) data were recorded using an Embla Model N7000 (MedCare, Bloomfield, CO) data acquisition unit outfitted with the software program Somnologica® 2.0. PSG records were exported from Somnologica into .edf format for EMG processing. The numerical computing program MATLAB (MathWorks© version 8.4.0 R2014b) was used to adapt the biosig toolbox version 2.93 [11] for conversion of all edf files into a .mat format, with a sampling rate of 200 Hz. Pre-processing included the manual removal of artifacts/spurious events from the final data set.

B. Brief Description of the Tool

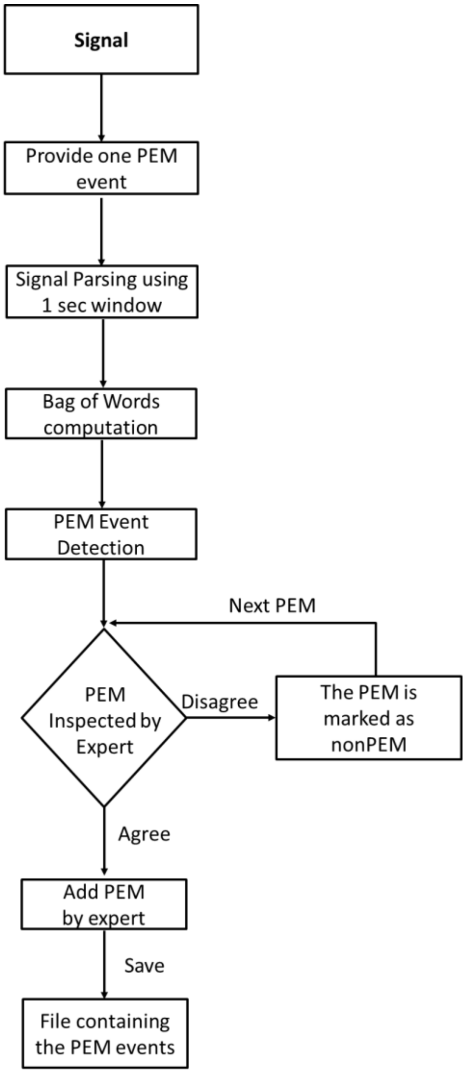


Figure 1. Flowchart of the PEM Scoring tool.

Our approach consists of the following steps:

i) The user must provide a PEM event,

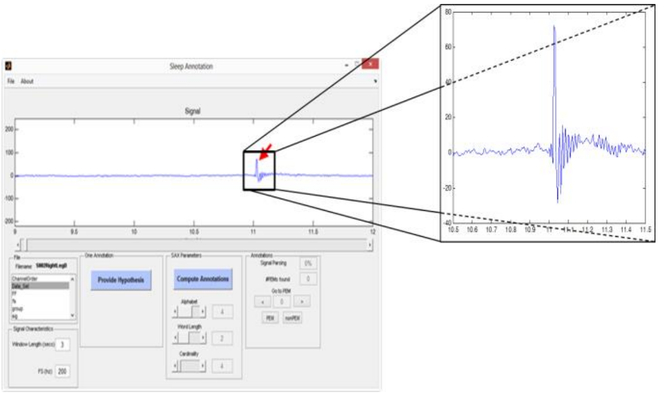
ii) the signal is parsed using a one second window,

iii) for each window the bag-of-words representation is computed with the help of Symbolic Aggregate approXimation (SAX) method

iv) the list of PEM events is computed and

v) Post-processing/Expert feedback to correct/remove misclassifications of PEM and Non-PEM events.

The flowchart of the proposed PEM annotator is depicted in Fig. 1, while a descriptive video is provided in (https://www.youtube.com/watch?v=kTTKq0xYoTM) in order to understand the use of the tool. Fig. 2a depicts a screenshot of the developed GUI, while Fig. 2b depicts the spectrum of the specific PEM event.


Sleep Annotation

Filename SM02Right.mat

Window Length (secs) 3

FS (Hz) 200

Alphabet 4

Word Length 2

Cardinality 4

Signal Parsing 0%

#PEMs found 0

Go to PEM 0

Figure 2. PEM annotator screenshot of the right leg EMG data from S002. The PEM event is demarcated by the red arrow.

C The Hypothesis & Signal Parsing

The cornerstone of the proposed method is that the user selects one PEM event. Pressing the “Provide Hypothesis” button, in the GUI annotator, the user can parse the signal quickly and find one PEM event using the mouse. The sample x_h which is closer to the point the user clicked is identified. Then a window of size of 200 points (one second long window) is defined centered at the point x_h. After the extraction of the window containing the PEM event, the bag-of-words representation for this window is computed.

D. Bag–of-Words Computation

One of the most well know methods to transform a time series into a symbolic string is the Symbolic Aggregate approXimation (SAX) method [12]. Given a real valued signal of N samples the SAX method produces a lower dimensional discrete representation of the original signal. However, two parameters need to be estimated: the size of the alphabet to use (i.e. A) and the size of the words to produce (i.e. w). The algorithm first normalizes the data and thus create a Piecewise Aggregate Approximation (PAA) representation [13], [14].

PAA transforms the original signal into a user defined number of segments. A signal X of length N can be represented in a n-dimensional space by a vector $\overline{X} = \overline{x}_1, \ldots, \overline{x}_n$. The i -th element of $\overline{X}$ is calculated as follows:

$$\overline{x}_i = \frac{n}{N} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \tag{1}$$

First the signal is normalized to have zero mean and standard deviation equal to zero. Then, equal sized windows (n) are used in order to reduce the time series from N dimensions to n dimensions. The mean value of the data falling within a frame is extracted and a vector of these values becomes the data-reduced representation. An example of the PAA of a time series is shown in Fig. 3.

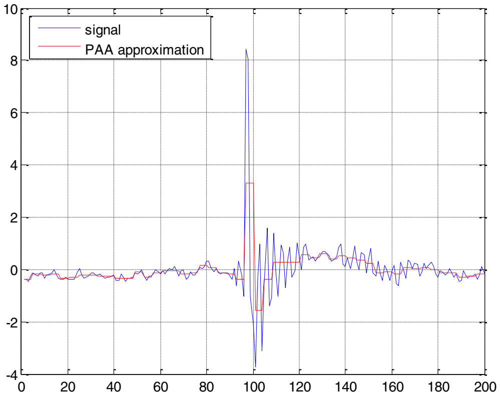


Figure 3. The PAA representation of a signal of length N=200 with n =50.

The normalization of the EMG signal results in an output that follows a Gaussian distribution from which breakpoints are selected such that equal sized areas under the Gaussian curve are obtained [15]. Finally, discrete signal representation is obtained using the following guidelines:

i) all the PAA coefficients which are smaller than the smallest break point are transformed to the symbol **a**

ii) all the PAA coefficients which are bigger or equal to the smallest break point and are smaller than the second smallest break point transformed to the symbol **b**

iii) all the PAA coefficients which are bigger to the second smallest break point and are smaller than the third smallest break point are transformed to the symbol **c**

iv) All the PAA coefficients which are bigger than or equal to the third smallest break point and are smaller than the fourth smallest break point transformed to the symbol **d**

v) Etc.

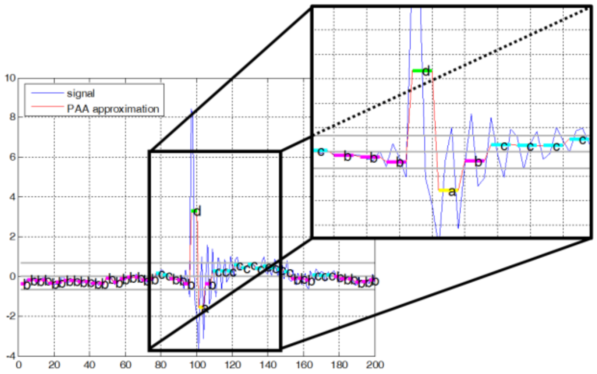


Figure 4. The discretization of the time series depicted in Fig. 3 using an alphabet of length 4 ($A = 4$).

Fig. 4 depicts the process described above for the signal displayed in Fig. 3, after the PAA transformation the parameters are set to $N = 200$ and $n = 50$, and the time series is mapped to the string:

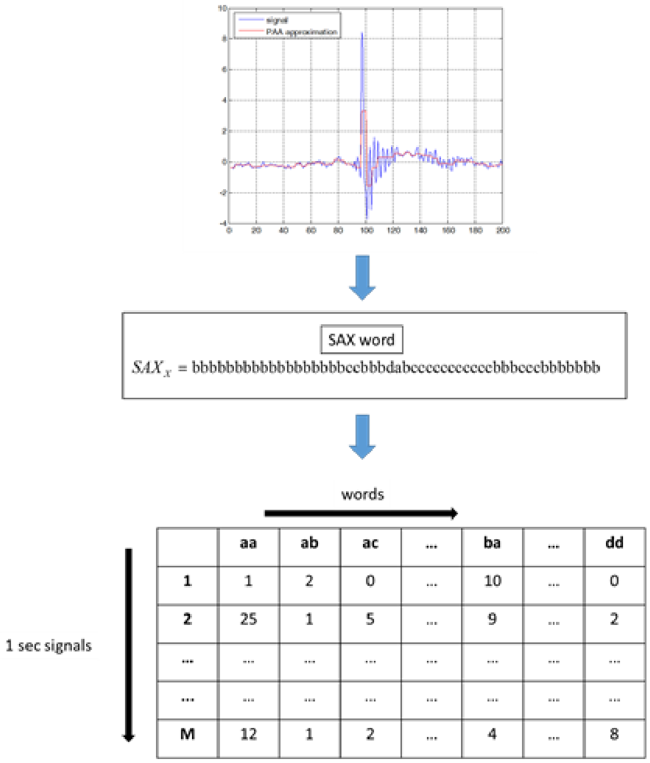
$$SAX_X = \text{bbbbbbbbbbbbbbbbbbccbbbdabcccccccccccbbbcccbbbbbbb} \tag{2}$$

After the computation of the SAX representation we have to transform the SAX string into the word-sequence matrix [16]. Given the size of the alphabet and the size of the words A and w respectively, the possible SAX words that can be produced are A^w . For example, for $A = 4$ and $w = 2$, our dictionary size is only 16. In [16] it has been shown experimentally that the choice of A does not critically affect the performance and a value of 3 or 4 works well for most time series.

The first attempt to map the sequence of letters into a bag–of-words representation would be to count the frequency of each letter. This can be extended by counting the frequency of words of length two etc. We can continue this process until the desired word length w . Thus from each one second window we compute the bag-of-words representation of each extracted SAX word. This is shown in Fig. 5.

C. PEM event detection

Only a single PEM event is required as the initial training set. During the detection of the PEM events, the whole set will be increased by more PEMs as the algorithm parses the signal. In the original proposal [8], each time series (each time segment in our case) would have been represented using SAX. However, because time invariance of PEM event occurrence within a time segment is important we adopt instead of the SAX representation the bag-of-words representation.



SAX_x = bbbbbbbbbbbbbbbbbbbbccbbbdabccccccccccccbbbccbbbbbbbb

	aa	ab	ac	...	ba	...	dd
1	1	2	0	...	10	...	0
2	25	1	5	...	9	...	2
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
M	12	1	2	...	4	...	8

Figure 5. The bag-of-words computation for a PEM signal.

Therefore, the single labelled PEM bag-of-words representation which is called hypothesis (Hs), will be considered as our model. If an unlabeled instance is very similar (considering a distance function) to Hs, its probability to be a PEM event will be high. However, the following method presents a significant disadvantage: without a stopping criterion the method will add all the unlabelled events to the PEM events set.

Let us define the Description Length (D_L) of a bag-of-words representation, $Words$, to be the total numbers of bits required for encoding:

$$D_L(Words) = L \cdot \log_2(\text{Card}) , \qquad (3)$$

where L is the length of the representation and $Card$ is the cardinality of the time series, we can define the Reduced Description Length of $Words$ given hypothesis Hs , $D_L(\text{Words} \mid \text{Hs})$, as the sum of the number of bits needed to encode $Words$ and the number of bits required for Hs itself, $D_L(Hs)$:

$$D_L(\text{Words}, \text{Hs}) = D_L(\text{Hs}) + D_L(\text{Words} \mid \text{Hs}) , \qquad (4)$$

The candidate events are labelled as PEM events as long as a compression reduction is achieved. Therefore, we encode the candidate event in terms of the hypothesis keeping the rest of the candidate events uncompressed as described here:

$$D_L(\text{Words}, \text{Hs}) = D_L(\text{Hs}) + D_L(\text{Words} \mid \text{Hs}) + D_L(\text{uncompressed}) , \qquad (5)$$

So as long we achieve compression, $D_L(Words) - D_L(Words \mid Hs) > 0$, we continue adding candidate events to the PEM events set.

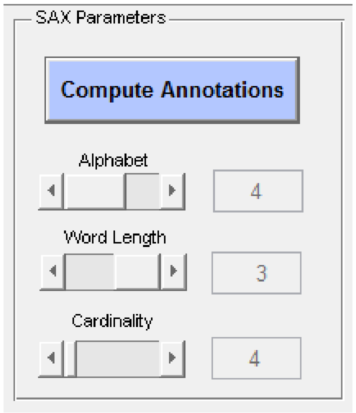
The following 3 step procedure summarizes our algorithm:

i) The nearest neighbor of any instance of our training set, which has not yet been labelled, is found.

ii) This unlabeled instance, is added to the training set.

iii) Repeat i) and ii) as long as $D_L(Words) - D_L(Words \mid Hs) > 0$.

The user/expert can set all the SAX parameters and the cardinality of the bag-of-words representation (alphabet, word length and cardinality) using the GUI as it is shown in Fig. 6.



Alphabet 4

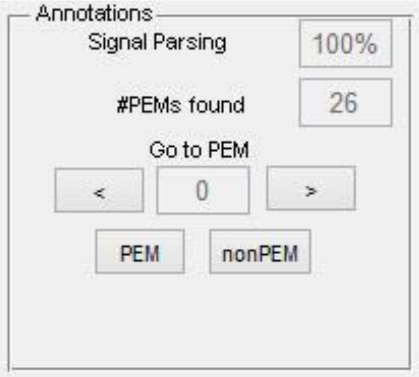
Word Length 3

Cardinality 4

Figure 6. Screen shot of toolbox, in annotation GUI, for setting SAX parameters .

F. Post-processing

The user/scorer can use the post processing module to access the PEM events visually and manually correct any automated misclassifications, from the previous step as shown in Fig. 7.



Signal Parsing 100%

#PEMs found 26

Go to PEM 0

Figure 7. The post processing module of the GUI.

III. Results

In order to conduct preliminary validation of the proposed approach for PEM annotation, the algorithm was tested on one EMG leg signal coming from a single subject. Table I contains a summary of PEM and non-PEM event classifications obtained using our approach, outlined in section II, where the golden-standard is visual/manual expert scoring [7].

TABLE I. CONFUSION MATRIX.

		True Class	
		PEM	***Non-PEM***
Predicted Class	PEM	18	30
	Non-PEM	20	801

Fig. 8 illustrates the reduction in the description length of PEM events coming from the aforementioned single subject's dataset.

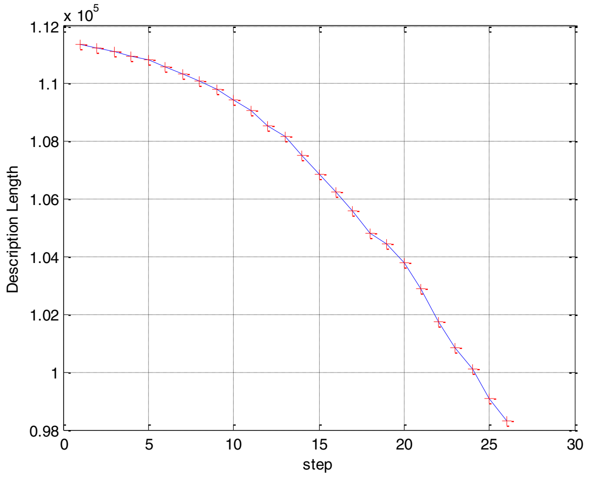


Figure 8. A plot of the reduction in description length with respect to using our algorithm for a single subject's PEM dataset.

IV. CONCLUSIONS

In this work we presented a semi supervised method for annotation of PEM using the bag-of-words representation of time series.

The novelty of the proposed approach stems from the fact that only one training sample is needed. Furthermore, the proposed approach relies on a modification of the approach proposed in [8]. In our case the MDL technique is not applied on the SAX representation, but on the bag-of-words representation which allows for time invariance in the occurrence of the PEM pattern.

Our preliminary results can be considered promising but further tuning of the method is needed using also data from more than one subjects. In future research we plan to test also an ensemble based approach allowing for the combination of different representations such as those relying on Fast Fourier Transform (FFT) [17], as well as accommodation for multiple training samples acting as "seeds" to the algorithm.

ACKNOWLEDGEMENT

This work was partially supported by the E.U. financed & national co-financed project "ICT System addressed to integrated logistic management and decision support for intermodal port and dry port facilities" Acronym: SAIL FP7-PEOPLE-IAPP-2009, Grant Agreement No. 251589.

REFERENCES

[1] D.L. Bliwise, L. He, F.P. Ansari and D.B.Rye, "Quantification of electromyographic activity during sleep: a phasic electromyographic metric," *Journal of Clinical Neurophysiology*, vol. 23, pp. 59–67, 2006.

[2] A. Iranzo, J. Santamaria, E. Tolosa, "The clinical and pathophysiological relevance of REM sleep behavior disorder in neurodegenerative diseases," *Sleep Medicine Reviews*, vol. 13, pp. 385-401, 2009.

[3] P. Karvelis, J. Fairley, G. Georgoulas, C.D. Stylios, D. Rye, and D. Bliwise, "Semi-Automated Annotation of Phasic Electromyographic Activity," *Artificial Intelligence: Methods and Applications Lecture Notes in Computer Science*, vol. 8445, pp. 532-543, 2014.

[4] J.A. Fairley, G. Georgoulas, N.A. Mehta, A.G. Gray, and D.L. Bliwise, "Computer detection approaches for the identification of phasic electromyographic (EMG) activity during human sleep," *Biomedical Signal Processing and Control*, vol. 7, no. 6, pp.606-615, 2012.

[5] J.A. Fairley, G. Georgoulas, O.L. Smart, G. Dimakopoulos, P. Karvelis, C.D. Stylios, D.B. Rye, and D.L. Bliwise, "Wavelet analysis for detection of phasic electromyographic activity in sleep: Influence of mother wavelet and dimensionality reduction," Computers in Biology and Medicine, 48, pp.77-84, 2014.

[6] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, vol. 2. MIT press, Cambridge, 2006.

[7] J. Fairley, P. Karvelis, G. Georgoulas, C. Stylios, D. Rye, and D. Bliwise, "Symbolic representation of human electromyograms for automated detection of phasic activity during sleep," *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 185-188, 2014.

[8] N. Begum, B. Hu, T. Rakthanmanon, and E. Keogh, "A Minimum Description Length Technique for Semi-Supervised Time Series Classification," Integration of Reusable Systems, *Special Issue in Advances in Intelligent and Soft Computing*, vol. 263, pp. 171-192, 2014.

[9] B. Hu, T. Rakthanmanon, Y. Hao, S. Evans, S. Lonardi, and E. Keogh, "Discovering the intrinsic cardinality and dimensionality of time series using MDL," *In: Proceedings of ICDM*, pp. 1086– 1091, 2011.

[10] T. Rakthanmanon, E. Keogh, S. Lonardi, and S. Evans, "Time series epenthesis: clustering time series streams requires ignoring some data," *In: Proceedings of ICDM*, 2011.

[11] C. Vidaurre, T.H. Sander, and A. Schlögl, "BioSig: the free and open source software library for biomedical signal processing," *Computational Intelligence and Neuroscience*, 2011.

[12] J. Lin, E. Keogh, W. Li, and S. Lonardi, "Experiencing SAX: A Novel Symbolic Representation of Time Series," *Data Mining and Knowledge Discovery Journal*, vol. 15, 2, pp. 107-144, 2007.

[13] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *J. Knowledge and Information Systems*, vol. 3, pp. 263-286, 2000.

[14] K. Yi, and C. Faloutsos, "Fast time sequence indexing for arbitrary Lp norms," *in Proc. of the 26th International Conference on Very Large Databases*, Cairo, Egypt, 2000.

[15] R.J. Larsen, and M.L Marx, *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, Englewood, Cliffs, N.J. 2nd Edition, 1986.

[16] J. Lin, E. Keogh, W. Li, and S. Lonardi, "Experiencing SAX: A Novel Symbolic Representation of Time Series," *Data Mining and Knowledge Discovery Journal*, vol. 15, 2, pp. 107-144, 2007.

[17] P. Schäfer, and M. Högqvist, "SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets," *In Proceedings of the 15th International Conference on Extending Database Technology*, pp. 516-527, ACM, 2012.