Understanding the Impact of Unexpected Cobot Movements on Human Stress Levels: A Time Series Classification Task

Lior Shilon[1,*,†], Vladyslav Shevtsov[1,*,†], Nadia Schillreff[2], Anne Rother[2], Shlomo Mark[1] and Myra Spiliopoulou[2]

[1]*Department of Software Engineering, Sami Shamoon College of Engineering, Ashdod, Israel*
[2]*Faculty of Computer Science, Otto-von-Guericke-University, Magdeburg, Germany*

Abstract

This study investigates the impact of unexpected stimuli on participants' stress levels during human-robot interactions (HRI). We designed an experiment, where a cobot performs writing tasks, as well as some unexpected actions, intended to surprise the human observer. During this experiment, we are continuously monitoring physiological responses through Electrodermal Activity (EDA) and Electrocardiogram (ECG) sensors. Our goal was to capture the relationship between surprising actions by the cobot and indicators of change in the human stress levels.

We employed time series classification techniques, namely ROCKET, Shapelet Transform and WEASEL, along with Support Vector Machine (SVM) analysis on temporal segments. This approach enabled us to develop a binary classification model on stress levels. Our research contributes to the design of AI systems that can detect changes in the stress levels of humans who interact with cobots that perform unexpected actions. Thus, we work towards the improvement of safety protocols in HRI.

Keywords

Human-Robot Interaction, Emotions in Human-Robot Interaction, Collaborative Robots, Time Series Classification

1. Introduction and Related Work

One of the major safety concerns in HRI is the potential for human-robot collisions, which can cause significant injuries, even cobots, when equipped with a sharp tool, can no longer be safe. Rapid, unexpected movements by robots can lead to involuntary human responses, such as startle or surprise, increasing the likelihood of such collisions. Recognizing changes in the stress levels, is crucial for improving safety and efficiency in HRI. Understanding these emotional and physiological responses is key to designing safer and more efficient robot trajectories.

Previous research has explored various aspects of HRI, particularly focusing on the detection and management of human emotions and physiological responses. Kirschner et al. [1] examined the effect of interactive user training on reducing human startle-surprise motion, highlighting

the importance of managing unexpected movements to prevent involuntary motions. This study demonstrated that hands-on training could significantly reduce involuntary motions, thereby enhancing safety in HRI. Our work is inspired by the need to recognize startle-surprise, even if it does not result in motion. As a first step in this direction, we conducted an experiment where the cobot is doing something surprising and we analyzed the stress levels of the human participant (one per experiment) before and after the surprising cobot motion.

Understanding the impact of unexpected stimuli on stress levels during human-robot interactions (HRI) is crucial for designing safer and more effective collaborative robots (cobots). Previous research has established that unexpected events can induce stress, which can be measured through physiological responses such as EDA and ECG signals [2, 3]. However, it is important to distinguish between general physiological arousal and specific stress responses. In this study, we assume that interruptions by the cobot lead to increased physiological arousal, which we interpret as stress within the experimental context. It is important to note that our findings do not map one-to-one to physiological responses and that ECG data had to be discarded due to poor contact quality.

Alonso-Martín et al. [4] presented a multi-modal emotion detection system that integrates voice and facial expression analysis to detect user emotions during HRI. The impact of surprising cobot motion on human stress levels is less investigated. Therefore, according to Klimek et al. [2], wearable EDA sensors are promising tools for detecting perceived stress, with field studies showing an average accuracy of 82.6% in predicting stress . The system, comprising GEVA (Gender and Emotion Voice Analysis) and GEFA (Gender and Emotion Facial Analysis), demonstrated high accuracy in emotion detection by combining the outputs of both channels through a decision rule. This multi-modal approach significantly improved detection rates compared to using individual channels. Inspired by this approach, we selected two physiological sensors, EDA and ECG, to capture human stress responses during HRI.

However, few studies have focused on the specific impact of surprise elements introduced by robots on human stress levels. Therefore, in this work we pursue following **Research question:** how to model and detect change in a human' stress levels, as caused by a cobot's surprising actions in a within-lab experimental setting? To address this question, we capture physiological data of the human, namely Electrodermal Activity (EDA) signal, monitor cobot activity, propose a workflow that partitions the human signal time series into segments containing human EDA and ECG recordings during the cobot's surprising action and segments of time where no such actions took place. Then, we deploy time series classification for segment separation and study the quality of the classifier. We postulate that high quality classification would indicate a change in the stress levels, which in turn would be associated to the only surprising event that occurs during the experiment, namely the cobot's surprising actions.

2. Experimentally Capturing Human Surprise in HRI

The primary aim of this experiment is to ascertain whether the proximity of a robotic arm engaged in writing, with the potential for unforeseen actions, exerts an influence on the stress levels of individuals. This influence will be quantified utilizing EDA and ECG sensors. The collected data are then used for stress level classification.

Designing HRI systems should balance surprise and predictability to maintain both engagement and trust. In our experiment, we achieved this balance by having the robot perform predictable writing tasks for most of the time, while incorporating short sequences with unexpected, i.e. surprising actions performed by the cobot. Such an action had the form of a sudden interruption of the cobot's writing activity and a movement of the cobot arm in another direction. Hereafter, we term this activity as 'surprising action (of the cobot). This design allowed us to measure the impact of unexpected actions on stress levels, while also considering the broader implications on engagement and trust as suggested by Law et al. [5].

2.1. Assumption of Stress Induction

We operate under the assumption that surprising actions of the cobot serve as indicators for time segments with change in stress levels. This assumption is based on established links between unexpected events and stress responses, as indicated by prior research Kirschner et al. [1]. Specifically, we expect that segments containing surprising actions will correspond to higher stress levels, as inferred from EDA signals. EDA has been shown to be a reliable indicator of stress, with wearable sensors demonstrating high accuracy in predicting perceived stress levels in various studies [2].

2.2. Measurement of Physiological Arousal

EDA is widely used to measure physiological arousal, which can indicate stress among other responses. Increased EDA reflects heightened sympathetic nervous system activity, often associated with stress. Studies have confirmed that EDA is an effective measure of stress, with reported accuracies between 42% and 100%, averaging 82.6% [2]. Although EDA measures general arousal and not exclusively stress, we label each segment containing a surprising action as 'positive' in the sense of being likely to contain an increase in stress level, and a segment containing no surprising action as 'negative'. Additionally, ECG has been shown to be a reliable tool for measuring stress [3]. However, we discarded ECG data altogether due to poor sensor contact, which compromised data quality.

2.3. Experiment Design

2.3.1. Setup

The experiments were conducted using the KUKA LBR iiwa robot [1]. We created and 3D-printed a tool that attaches to the robot's flange and holds the writing implement (pen) used by the robot to write. We then programmed the robot to transcribe the input sentences onto a sheet of paper placed on a table in front of the human participant.

2.3.2. Types of surprising actions

To model surprise, five different types of actions were designed:

---

[1]https://www.kuka.com/en-de/products/robot-systems/industrial-robots/lbr-iiwa   Kuka   Robotics. http://www.kuka.com "Kuka" is a registered trademark of KUKA Deutschland GmbH.

1. The robot stops writing mid-sentence.
2. The robot returns to the starting position.
3. The robot makes a sudden movement towards the participant.
4. Small spline movement. (A small movement in space of the robotic arm close the base of the arm.)
5. Big spline movement. (A bigger movement that outreaches more towards the participants.)

The first two actions were designed by Participant 1, the third by Participant 2, and the last two by a 3rd party that didn't have the experiment run on them. actions 1 and 2 are non threatening movements (the robot is not moving towards the human), while actions 3, 4, and 5 differ in the amount of movement (for example, action 5 utilizes more workspace for the movement). We opted for this approach to minimize exposure to the surprises as much as possible, given our dual roles as both the experiment's programmers and participants.

To measure the impact of the Cobot's actions on human stress levels, Electrodermal Activity (EDA) and Electrocardiogram (ECG) sensors are utilized. These sensors capture physiological responses indicative of stress and arousal levels:

- **Electrodermal Activity (EDA)**: Measures changes in skin conductance, reflecting sympathetic nervous system activity. Increased EDA suggests heightened arousal or stress.
- **Electrocardiogram (ECG)**: Records the electrical activity of the heart, providing insights into heart rate variability (HRV) and cardiac responses to stressors.

2.3.3. Hypothesis

We hypothesize that the EDA/ECG levels of participants will differ between the segments that contain a surprising actions of the cobot and the segments that do not.

2.3.4. Procedure in the Experiment

Pre-Experiment Setup: (1) Position the robot in front of the table and secure it to prevent unwanted movements. (2) Place and secure the paper on the table using tape. (3) Attach EDA and ECG sensors to participants according to manufacturer instructions. (4) Set up cameras to record the experiment.

Note that the camera is used to identify the exact time points where each surprising action started and ended, so that the segments with the physiological signal can be marked accordingly.

Experimental Phase: (1) The robotic arm writes a text on a surface; this is visible to the participant of the experiment. (2) Each participant undergoes two experiments/sessions, with up to 6 surprising actions per experiment.

Each surprising action has a 0-37% chance of occurring randomly during the writing phase, with a minimum of 5-10 movements between surprising actions. surprising actions that haven't happened yet will take precedent over ones that did.

Data Collection: (1) Record EDA and ECG readings throughout the experiment. (2) Prompt participants to rate their stress levels after each session. (3) Record any additional observations or qualitative data relevant to stress levels.

Post-Experiment tasks: (1) Remove the sensors from the participants' hands. (2) Log the sensor data. (3) Shut down the robot.

2.4. Building Time Segments for Classification on Surprise

We segmented the recorded physiological data (EDA and ECG) in 1-second intervals. We also used video recordings, which helped us to identify and mark in the time series the precise starting point and ending point of each surprising action. For EDA data, we use a minimum segmentation of 20 units, and for ECG data, 25 units, accommodating the longest observed interruption. Then, we assigned labels to the segments, as explained in subsection 2.2, where 'positive' corresponds to class 1 and 'negative' to class 0. We used these segments to train and test and time series classifier.

We show the concrete EDA time series of Participant 1 in the two plots of Figure 1, and alike the EDA time series of Participant 2 in the two plots of Figure 2. We marked the 0-segments as red and the 1-segments as green. The starting time point in each time series is the moment where the participant taps at the sensor - and then the recording starts.
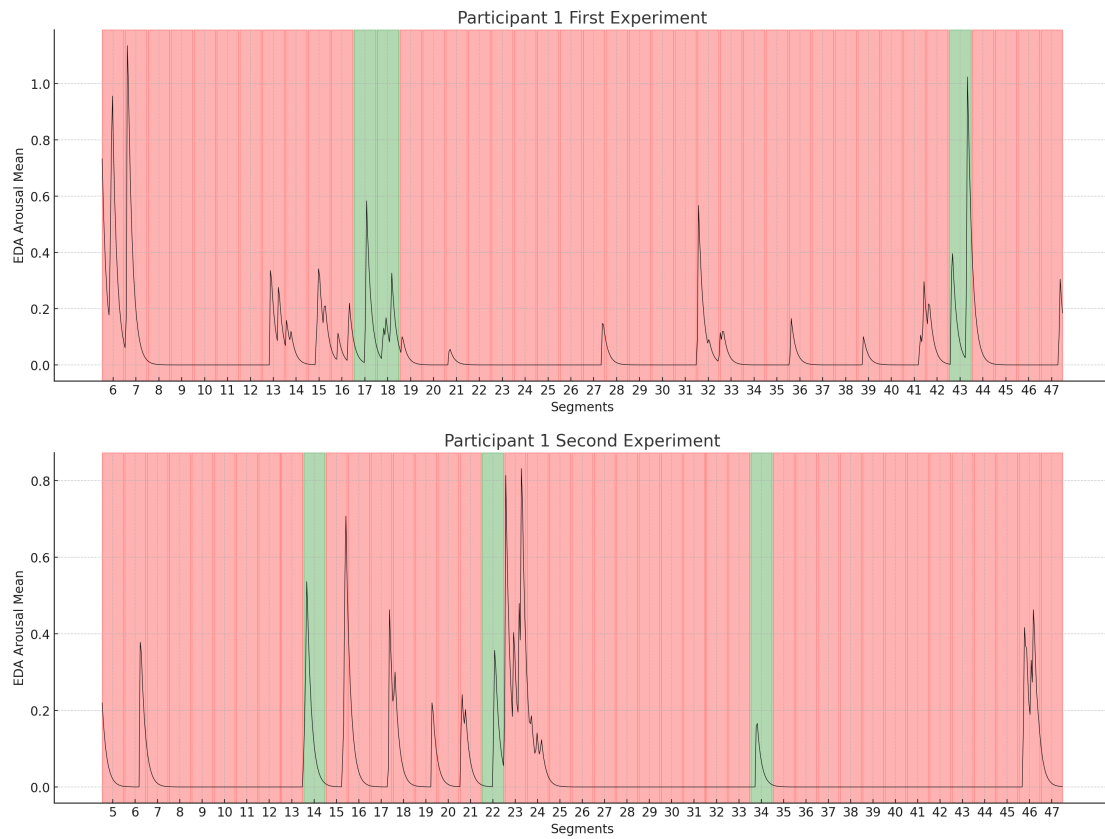

Figure 1: The EDA time series of participant 1, first then second experiment: segments with surprising actions (class 1) are marked in green, segments without (class 0) are marked in red

Figure 2: The EDA time series of participant 2, first then second experiment: segments with surprising actions (class 1) are marked in green, segments without (class 0) are marked in red
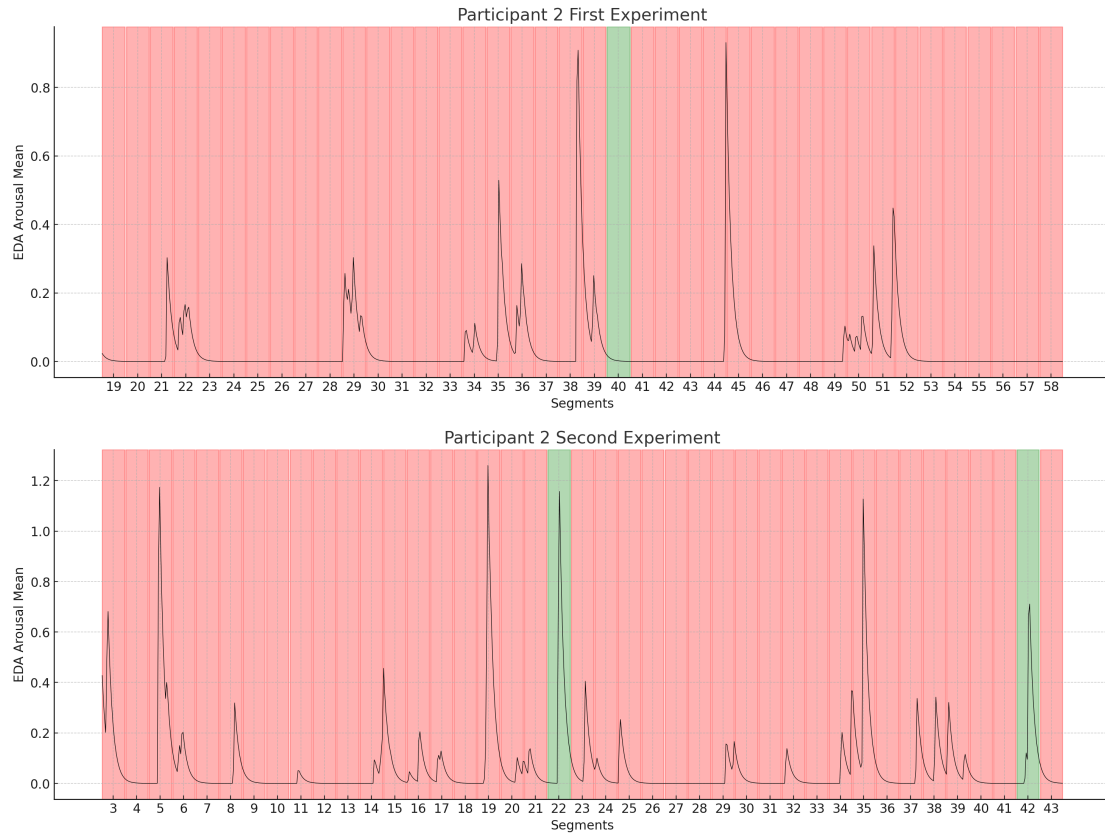
As can be seen in the two figures, and particularly in Figure 2, some peaks at the stress levels are *after* or even *before* the time interval where the surprising action took place: for a case with a peak well before the surprising action, refer to time point 36 of the first experiment of participant 2. Moreover, there are segments with peaks, although there is no surprising action. This can be attributed to some external event that caught the participant's attention, or to some misalignment between video recording and EDA recording. These artifacts were ignored during time series classification, but it is evident that they made the classification task much more difficult.

Notwithstanding the aforementioned artifacts, Table 1 depicts the number of segments per participant and experiment; it is evident that the class distribution is skewed, class 1 is underrepresented. Note that we show only the EDA time series, because the ECG signals were not properly captured and had to be discarded.

2.5. Preparing the Data for Time Series Classification Algorithms

Building on the segmented data from the previous subsection, we carried out additional steps to make the data suitable for classification.

		# positive (1) segments	# negative (0) segments	Total
Participant 1	Experiment 1	3	39	42
	Experiment 2	3	40	43
Participant 2	Experiment 1	1	39	40
	Experiment 2	2	39	41

Table 1
Distribution of positive and negative segments per participant and experiment

We computed statistical features for each segment for the SVM classifiers, including the mean, standard deviation, and median. We also tabulated the count of positive (stress response) and negative (no stress response) instances.

For classification, we used several algorithms including ROCKET, Shapelet Transform, and WEASEL classifiers. Additionally, we trained three SVM classifiers with different kernels (linear, RBF, and sigmoid) on both 1st participant's and 2nd participant's experimental data, testing them across each other's datasets using the extracted features.

2.6. Time Series Classification Models

2.6.1. ROCKET

According to Dempster et al. [6], "ROCKET transforms time series using a large number of random convolutional kernels, i.e., kernels with random length, bias, weights, dilation, and padding." Essentially, ROCKET breaks the time series data into small windows, identifying the highest point (global max) and calculating the proportion of positive values (ppv) [6]. In our context, a time series is a segment.

2.6.2. WEASEL

The WEASEL algorithm, as described by Schäfer and Leser [7], uses a sliding window to extract sub-sequences, then transforms each one into a feature vector by applying Symbolic Fourier Approximation (SFA) Schäfer and Leser [7]. As with ROCKET, the inputs to WEASEL are the segments.

2.6.3. Shapelet Transform

The Shapelet Transform method by Bostrom and Bagnall [8] and Hills et al. [9] selects *randomly* a subset of sub-sequences from the time series data as potential shapelets, thus avoiding exhaustive search for shapelets of all possible lengths. These potential shapelets are compared to all possible sub-sequences in the data [8, 9]. For our data, these are sub-sequences of each segment, keeping in mind that segments may vary in length. Finally, features are extracted based on the similarity between each time series and the selected shapelets [8, 9].

2.7. Training and Testing

We constructed Time Series Classifier (TSC) models using the `sktime` library, specifically employing the ROCKET Classifier (multirocket), WEASEL, and Shapelet Transform classifiers.

A total of nine time series models were trained. The models included: one trained solely on 1st participant's first experiment data; one trained on both 1st participant's first and second experiment data; and one trained on 1st participant's first and second experiment data in addition to 2nd participant's first experiment data.

All models were evaluated using data from 2nd participant's second experiment, which was selected due to its higher frequency of interruptions.

3. Results

Our results are depicted on Table 2, where we evaluate on precision, recall and F1 towards the positive class 1. We didn't add SVM to the table because every SVM model we trained predicted only 0 for everything. It is evident that no model could cope with the extreme skew in favour of class 0: ROCKET focussed one class 1 when training on participant 1, Shapelet Generation did so only when using all data from participant 1, while WEASEL concentrated on class 0 throughout. The usage of the data of the second participant did not contribute to model quality. This indicates that oversampling of class 1 or under-sampling of class 0 or both should be done before learning on such skewed data.

Classifier	Scenario	Precision	Recall	F1
ROCKET	Experiment 1 of Participant 1	0.29	1.00	0.44
	Experiments 1 & 2 of Participant 1	0.20	1.00	0.33
	Experiments 1 & 2 of Participant 1, Experiment 1 of Participant 2	0.00	0.00	0.00
WEASEL	Experiment 1 of Participant 1	0.00	0.00	0.00
	Experiments 1 & 2 of Participant 1	0.00	0.00	0.00
	Experiments 1 & 2 of Participant 1, Experiment 1 of Participant 2	0.00	0.00	0.00
Shapelet Generation	Experiment 1 of Participant 1	0.00	0.00	0.00
	Experiments 1 & 2 of Participant 1	0.50	1.00	0.67
	Experiments 1 & 2 of Participant 1, Experiment 1 of Participant 2	0.00	0.00	0.00

Table 2
Performance metrics for each scenario and model

4. Conclusion

We presented an experiment design to measure change in stress levels due to surprising cobot actions, and a framework for the analysis of the experimental data. We measured electrodermal activity (EDA) of the participants, and we partitioned the EDA time series into segments that correspond or do not correspond to surprising actions (positive/negative). The analysis

encompassed the separation between segments of the two classes with time series classifiers. For this, we created a workflow that allowed the comparison of models build by different state of the art time series classification algorithms. For this workflow, we also used several evaluation criteria, namely precision, recall and F1 for each of the classes, albeit we only reported for the rare class here.

The purpose of the analysis was to investigate to what extend it is possible to distinguish between the two types of segments by considering only EDA as stress indicator. The hypothesis could not be tested because the number of positive segments was too small. Furthermore, the presence of peaks in the stress levels well after or even well before the actual surprising action indicates either that the participant was distracted or that the time interval of the surprising action could not be reproduced properly from the video; this corresponds to a misalignment between video recording and time series of body signal. This affected the classification negatively. Moreover, the experimental data were reduced because the ECG signal recordings indicated a sensor fault and had to be ignored. A further limitation was the small number of experiment participants. Nonetheless, the data skew was the primary limiting factor, and should be mitigated in future work through oversampling of the positive class and/or under-sampling of the negative class. While under-sampling is straightforward, oversampling of temporal segments may demand the generation of synthetic time series, which is itself challenging in the case of so few positive samples.

We suspect that the duration of the surprising actions also affected the EDA recordings. In particular, surprising actions of long duration translate into long segments, where a short peak in the stress level might be misinterpreted as noise. Hence, future experiments should achieve a balance between duration of the surprising actions and number of such actions. Finally, a larger number of participants should be recruited. Then, it will become interesting to investigate to what extend participant-specific models can be augmented with data or models of other participants.

Acknowledgments

References

[1] R. Kirschner, L. Burr, P. M., H. Mayer, S. Abdolshah, S. Haddadin, Involuntary motion in human-robot interaction: Effect of interactive user training on the occurrence of human startle-surprise motion, in: 2021 IEEE International Conference on Intelligence and Safety for Robotics, 2021.

[2] A. Klimek, I. Mannheim, G. Schouten, E. J. M. Wouters, M. W. H. Peeters, Wearables

measuring electrodermal activity to assess perceived stress in care: a scoping review, Acta Neuropsychiatrica (2023) 1–11. doi:10.1017/neu.2023.19.

[3] M. Naeem, S. A. Fawzi, H. Anwar, A. S. Malek, Wearable ecg systems for accurate mental stress detection: a scoping review, Journal of Public Health (2023). URL: https://doi.org/10.1007/s10389-023-02099-6. doi:10.1007/s10389-023-02099-6.

[4] F. Alonso-Martín, M. Malfaz, J. Sequeira, J. F. Gorostiza, M. A. Salichs, A multimodal emotion detection system during human–robot interaction, Sensors 13 (2013) 15549–15581. URL: https://www.mdpi.com/1424-8220/13/11/15549. doi:10.3390/s131115549.

[5] E. Law, V. Cai, Q. F. Liu, S. Sasy, J. Goh, A. Blidaru, D. Kulić, A wizard-of-oz study of curiosity in human-robot interaction, in: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017, pp. 607–614. doi:10.1109/ROMAN.2017.8172365.

[6] A. Dempster, F. Petitjean, G. I. Webb, Rocket: exceptionally fast and accurate time series classification using random convolutional kernels, Data Mining and Knowledge Discovery 34 (2020) 1454–1495. URL: https://doi.org/10.1007/s10618-020-00701-z. doi:10.1007/s10618-020-00701-z.

[7] P. Schäfer, U. Leser, Fast and accurate time series classification with weasel, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, ACM, 2017. URL: http://dx.doi.org/10.1145/3132847.3132980. doi:10.1145/3132847.3132980.

[8] A. G. Bostrom, A. J. Bagnall, Binary shapelet transform for multiclass time series classification, Trans. Large Scale Data Knowl. Centered Syst. 32 (2015) 24–46. URL: https://api.semanticscholar.org/CorpusID:265852316.

[9] J. Hills, J. Lines, E. Baranauskas, J. Mapp, A. Bagnall, Classification of time series by shapelet transformation, Data Mining and Knowledge Discovery 28 (2013). doi:10.1007/s10618-013-0322-1.