Symbolic Representation of Multivariate Time Series Signals in Sport Activity Classification

Matarmaa Jarno

Granted by
Ministry of Science and Higher Education of the Russian Federation
(Ural Federal Univeristy Program of Development within the Priority – 2030 Program)
is gratefully acknowledged

Institution
Ministry of Education and Science of the Russian Federation
Federal State Autonomous Educational Institution of Higher Education
«Ural Federal University named after the first President of Russia B. N. Yeltsin»
Engineering School of Information Technologies, Telecommunications and Control Systems

Abstract

This study introduces a new multivariate time series (MTS) sport activity dataset consisting of five categories, walking, running, biking, skiing, and roller skiing. The original data of 228 activities have been recorded by an individual athlete for a 16 months time period in uncontrolled environments using two types of sport watches. The dataset consists of three-dimensional multivariate time series features such as heart rate, speed, and altitude, which are popular and pure sensor based attributes for endurance outdoor sport activities. The pre-processed signals are split into 69 seconds equal length segments and several segments from each single activity have been gathered in order to conduct data augmentation because of a relatively small dataset size. The MTS classifier called WEASEL+MUSE was applied to the dataset in order to discriminate categories based on the time series characteristics of the signals. WEASEL+MUSE implements word extraction from the signal -method by building a multivariate feature vector using a sliding-window approach applied to each dimension of the MTS, then extracts discrete features per window and dimension. WEASEL+MUSE, developed in 2016, have been one of the most successful approaches in multivariate time series classification. The classification results was analyzed using several popular quality metrics and tools such as ROC curve. In addition, an early time series classification (eTSC) algorithm called TEASER was applied to determine how much data will be sufficient to find a balance in accuracy and computation time tradeoff. According to the results, dataset integrity is generally good and sport activities were classified fast and accurate, up to 93,0%. Most of the problems were identified between very similar activities which can be practically impossible to discriminate perfectly by any algorithm or dataset. Signal length analysis indicated that 33% of the data will provide relatively good results, 85,6% accuracy in the test data.

Keywords: multivariate, time series classification, sport activity dataset, word extraction, symbolic Fourier approximation, MTS, SFA, WEASEL, MUSE

1. INTRODUCTION

1.1 Problem description

Nowadays sport watches are collecting and storing an enormous amount of data from a wide variety of sport activities. Sport watches, as well as many kinds of smart watches and smartphones which are able to record data from sport activities using multiple different type of sensors, are using manually selectable sport profiles when starting a new activity recording in order to label and classify them. For each sport different types of data attributes are tracked during training which will be defined by sport profile selected before starting to record the activity. Thus, there are many problems which possibly cause false sport activity labelling. The most common situations might be that 1) humans can select wrong sport profile accidentally, 2) smartwatch or recording device may not have actual sport profile, or 3) the wrong sport profile is chosen intentionally or due to indifference. This wrong labelling or unreliable human made classifying is problematic since sport activity tracking platforms have also become the one form of social media where people can discuss and compare their sport activities among each other. But also, wrongly labelled activities might cause distortion in general personal data statistics, and therefore misleading guidance from smartwatch to user since modern smartwatches are highly interactive. This study introduces a time series based method for retrospective supervised sport activity classification (SAC) in order to correct mislabeled data afterwards for any sport activity dataset of a single person. The base idea is to train classification model separately for a single person, leading to an enhanced classification accuracy since interpersonal differences do not need to be considered.

1.2 Literature overview

The research field regarding and associated with introduced problem is called Human Activity Recognition (HAR). During the past two decades it has been widely investigated study field due to fast increasement in popularity and development of activity bracelets, smartwatches and smartphones providing sensors to measure appropriate data through inertial sensors. Advance of deep learning and machine learning algorithms has allowed researchers to use HAR in various domains including sports, health, and well-being applications. [1] The goal of HAR is to recognize human activities in controlled and uncontrolled settings.

Lara and Labrador (2013) implemented a comprehensive and unquestionably high-quality overview in HAR studies, and it is honor to point out some great aspects from their study regarding a personalized sport activity classification, despite the fact that many new algorithms have been developed after their study. They identified an open debate on the design of any activity recognition model. Since according to some authors, people perform activities in a different manner as they differ on age, gender, weight, and so on, a specific recognition model should be built for each individual. [2] This implies that the system should be retrained for each new user [3]. However, many studies rather emphasize the need of a monolithic recognition model, flexible enough to work with different users [4]. Consequently, two types of analyses have been proposed to evaluate

activity recognition systems: subject-dependent and subject-independent evaluations [5]. In the first one, a classifier is trained and tested for each individual with his own data and the average accuracy for all subjects is computed. In the second one, only one classifier is built for all individuals using cross validation or leave-one-individual-out analysis. Lara and Labrador highlighted that in some cases it would not be convenient to train the system for each new user, especially when 1) there are too many activities, 2) some activities are not desirable for the subject to carry out, or 3) the subject would not cooperate with the data collection process. Furthermore, older people would surely walk quite differently than a young child and is therefore challenging a single model to recognize activities regardless of the subject's characteristics. They suggested that a solution to the dichotomy of the monolithic and particular recognition model can be addressed by creating groups of users with similar characteristics. [2] However, in HAR studies we are often limited by the availability of such datasets.

Since in many studies accuracies up to 99% have been achieved in a controlled data collection experiment one may question if the environment of data collection has insufficient versatility or unilateral too unrealistic conditions ignoring many factors when performing activities in real environments. As in some studies have been noticed also remarkable accuracy drops for activities from controlled data collection experiment to uncontrolled non-laboratory natural environment [6], it might be reasonable to conduct classification tasks in more challenging datasets to ensure, that all environmental variables will be considered. This study addresses these aforementioned issues by introducing and publishing a new unique dataset, recorded in a genuine environment without the owner of the data being aware of the possibility to utilize it in research.

1.3 Study restrictions

In this study several limiting factors are acknowledged. Probably the most important is the lack of data from a number of athletes in order to investigate interpersonal differences in classification accuracy, and thus generalize achieved results. Another important question could be that what is the significance of conducting SAC as a time series classification instead of using simply extracted summary features from the signals and then apply traditional machine learning model or neural network. However, the study question is set as follows: can we add some value for classification considering intercorrelation of column values, and therefore to solve the task using multivariate time series classification approach in uncontrolled environment generated dataset of individual. The goal is not to find the best possible method for sport activity classification.

Since the dataset comes from an individual athlete interpersonal differences are not considered in the study. Many difficulties can be identified when inspecting possible weaknesses, and yet some of them will remain unidentified. For example, as we record sensor data like heart rate, that may have very unique behavior among sports depending on the person the model must definitely be trained for each person separately. That has been acknowledged when conducting this study. Lastly, the applied dataset can be classified with order of magnitude faster and better accuracy by extracting basic summary features from the time series data and thereafter implement classification as a standard CML task. However, this study try to answer another question rather than

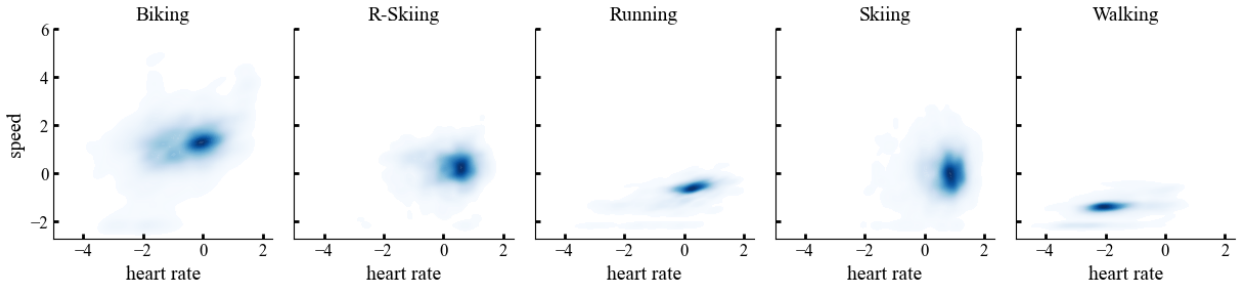

Figure 2.1: Kernel Density Estimation (KDE) plots by category for heart rate and speed features. Skiing and R-Skiing have quite similar kernel density whereas others are more distinctive.

find the best method to classify the sport activities of the dataset. Thus, the question is set as follows: is it possible or further reasonable to classify outdoor sport activities using three dimensional signals including heart rate, speed and altitude by extracting time series features from them? That question will be answered in this study.

2. DATASET

Description

In this study we introduce a new multivariate time series dataset including 228 outdoor sport activities in five categories, *walking*, *running*, *biking*, *skiing* (cross country skate skiing), and *roller skiing* (R-Skiing), recorded in 16 month time period by an individual non-competitive male athlete at the age of 32 yeas. Dataset includes three dimensions or attributes which are measured by three different type of sensors: heart rate, geolocation and barometer sensors. For extracted feature format, signals are transformed into heart rate, speed, and altitude, respectively. Dataset has been pre-processed and cleaned according to a particular criterion based on the subject domain knowledge, and thereafter segmented and standardized. Sport activities with missing sensor data or too short in length have been simply dropped out since these are crucial factors in order to conduct credible high quality classification in relatively small size dataset. In the segmentation a simple procedure was conducted: from the beginning of the original sport activity – starting from the point of 100 seconds – 69 seconds intervals were selected. Then, as a compensation for a small dataset size, data augmentation has been implemented by picking 5 consecutive 69 seconds intervals from the same activity. Thus, as an end result, we have a three-dimensional dataset in the size of (1140, 69, 3). This data is stored in separate files for each of the three attribute, *hear rate*, *speed*, and *altitude*. Before applying this data to the sktime classifier, we have to combine these files to a nested data structure as introduced in the study of Löning et al. (2019) [7]

Category data visualizations

Since dataset has three dimensions, one of them will be ignored in visualization, namely altitude which does not have such an unambiguous relationship with speed and hear rate. And because we are familiar with the fact that speed and heart rate features have a quite strong positive

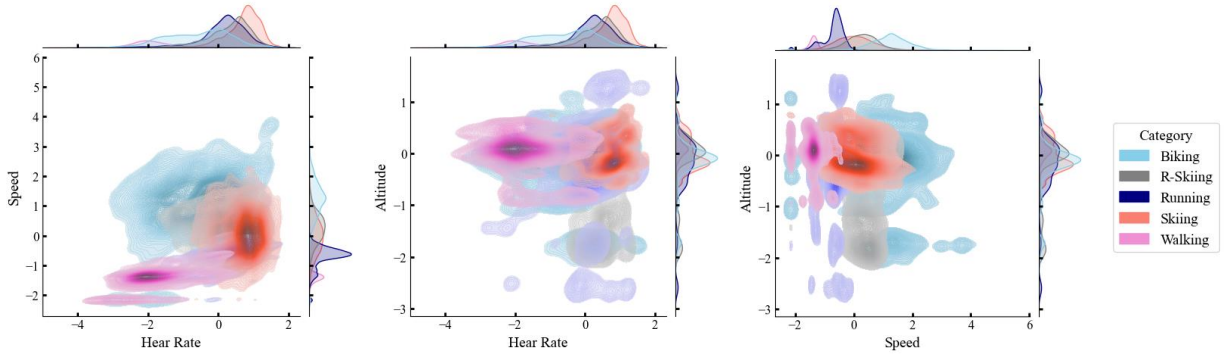

Figure 2.2: Kernel Density Estimation (KDE) graphics by category for heart rate and speed features. In addition to bivariate KDE, value distribution is demonstrated using univariate KDE as a joint plot in the same figure.

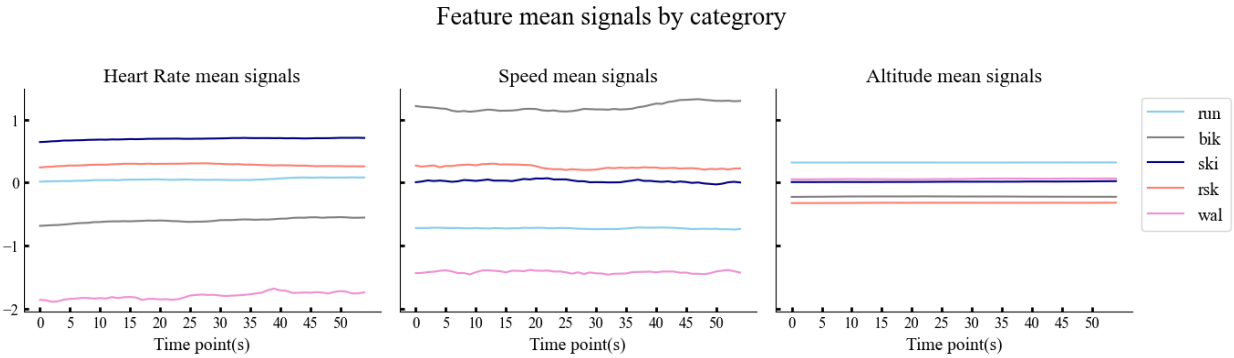

Figure 2.3: Feature mean signals by categories for Heart Rate, Speed, and Altitude. As a conclusion, the categories of the data are quite clearly distinguishable, and lines do not have intersection points.

correlation which behaves differently among categories, we use these ones in reduced two dimensional data space in order to visualize data descriptively. Figure serie 2.1 shows how the speed-hear rate value points are located among categories. From these pictures can be seen, that Skiing and R-Skiing patterns have quite similar shapes while other categories are more discriminative.

When kernel density graphics of categories are combined into the same figure, intersection areas can be observed (fig. 2.2). Here altitude feature is also visualized as it becomes convenient. Density values are normalized by category and thus considers unbalanced number of instances among categories. As a quick pre-analysis can be stated that according to heart rate and speed feature relation Skiing and R-Skiing are almost impossible to discriminate, but they become a little bit more distinctive when inspecting a relation of these features with altitude dimension.

Signal characteristics

Time series of the dataset may have varied and divergent shapes even among the same category but when we generate mean signals by categories they can be presented as lines which does not intersect each others at any point in 55 seconds signals, as shown in figure 2.3. It might probably indicate that we may use even a single number as an eigenvalue for each category to describe the whole segment and compare them without picking any certain point from the segment.

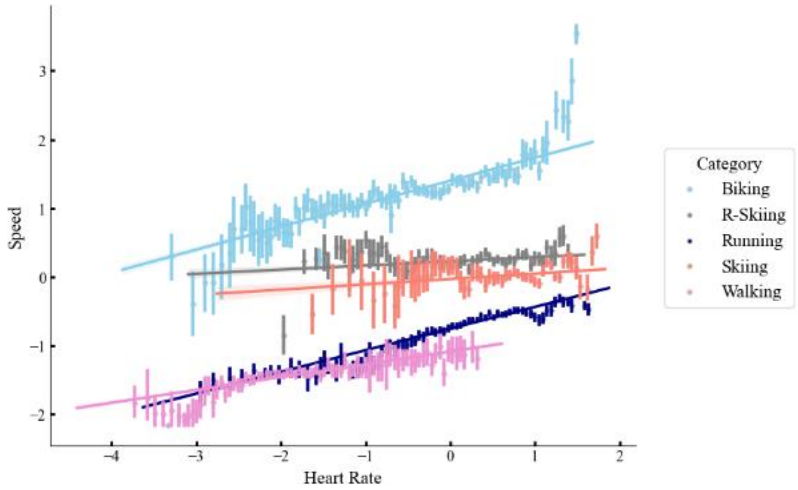

Figure 2.4: Scatter plot with regression lines of Speed and Heart Rate features. For all the categories speed and heart rate have positive correlation: Running and Biking having considerably greater correlation than others whereas similar Skiing and R-Skiing have very similar and only moderate or very low correlation.

Thus, it is quite obvious that decision three based model with using only mean values might work very well in classification of these signals.

Yet another way to describe discriminative characteristics of the categories is scatter plots with regression lines in heart rate and speed features. However, domain knowledge of the relation of altitude feature to speed and heart rate does not provide any reason for inspecting its correlation, despite the fact that in certain circumstances it could produce discriminatory properties. Inspecting the figure 2.4 can be observed that in all the categories speed and heart rate have positive correlation but Running and Biking have considerably greater correlation than others whereas similar Skiing and R-Skiing have very similar and only moderate or very low correlation. For the simplicity figure applies binning strategy with a value of 300 to make a better contrast between categories.

Differential features

Also, differential features of hear rate speed and altitude were investigated. In the analysis were observed decreasing trend between Speed and Altitude derivatives among all the activity data, which means, that when an athlete goes uphill speed decreases (speed derivative is negative). However, the trend is quite moderate because speed can decrease for several other kind of reasons which makes interpretations difficult. Further, according to the differential feature analysis of heart rate and altitude there is slight positive correlationn. Positive correlation is expected, but as for speed, heart rate can increase for many other kinds of reasons which makes interpretations difficult. Therefore, analysis should consider only those segments when absolute altitude derivative has greater value than some certain appropriate selected threshold value to reject steady altitude segments and make data more interpretable. That requires a massive signal processing procedure before classification, and it would form a high-level case optimized application, and therefore is out of the target of this study. However, these aspects are mentioned in order to understand and interpret possible outcomes better.

Data setup for classification

In the classification, we will reserve 20% of the data for model validation. That means there are 912 instances in the train data and 228 in the test data. Therefore, one instance has $\frac{100}{228} \approx$ 0,44% weight in the results. This information provides a perspective for interpretation of the results and should be taken into account.

3. CLASSIFICATION METHOD

3.1 MUSE algorithm

For multivariate TSC we use Multivariate Unsupervised Symbols and Derivatives (MUSE) algorithm, also known as WEASEL+MUSE, which is implementation of multivariate version of WEASEL (Word Extraction for time Series Classification) and in this study referred as just MUSE. MUSE is a multivariate dictionary classifier that builds a bag-of-patterns (BOP) using Symbolic Fourier Approximation (SFA) for different window lengths and learns a logistic regression classifier on this bag. [8]

The novelty of MUSE lies in its specific way of extracting and filtering multivariate features from MTS by encoding context information into each feature. It uses statistical feature selection, derivatives, variable window lengths, bigrams, and a symbolic representation for generating discriminative words. MUSE provides tolerance to noise due to use of the truncated Fourier transform, phase invariance, and superfluous data/dimensions. Thereby, MUSE assigns high weights to characteristic, local and global substructures along dimensions of a multivariate time series. It has achieved good results even for small-sized datasets, where deep learning based approaches typically tend to perform poorly. When looking into application domains, it is best for sensor readings, followed by speech, motion and handwriting recognition tasks. [8]

3.2 Word Extraction (SFA)

The MUSE algorithm is not an actual classification algorithm but rather a pipeline or interface that encapsulates complex time series transformation methods with a traditional CML algorithm which is logistic regression. Therefore, it is reasonable to make clear also underlying data processing principles. As the introduced new dataset has become through a quite massive preprocessing and structuring process, the MUSE algorithm further continues to transform data to a proper format. As mentioned, it uses SFA transformation algorithm, which consist of preprocessing and transformation phases. Preprocessing includes Discrete Fourier Transform (DFT) approximation and Multiple Coefficient Binning (MCB) discretization. Approximation algorithms try to capture the most important information from time series, and thus they can be seen as simple feature extraction algorithms. Binning continuous data into intervals can be seen as an approximation that reduces noise and captures the trend of a time series. The MBC bins continuous time series into intervals, transforming each time point of all the time series (a sequence of floats) into a sequence of symbols, usually letters. Controversary to the Symbolic Aggregate Approximation

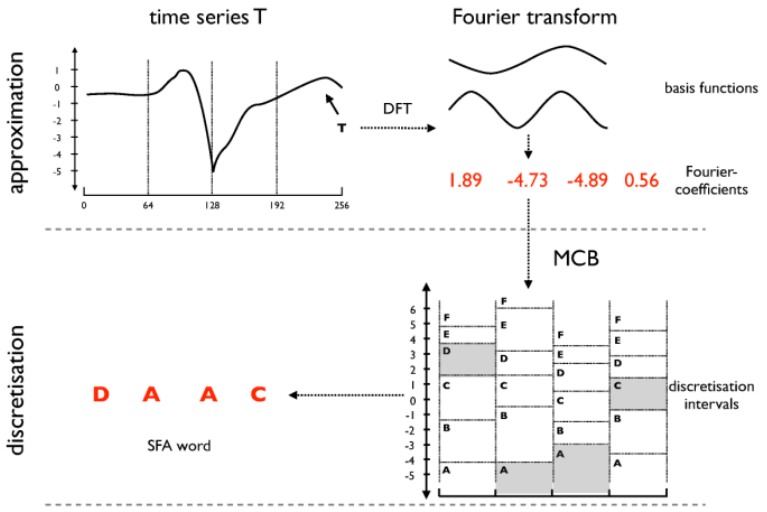

Figure 3.2A: SFA process: Time series approximation (DFT), discretization (MCB), and transformation (SFA). [9]

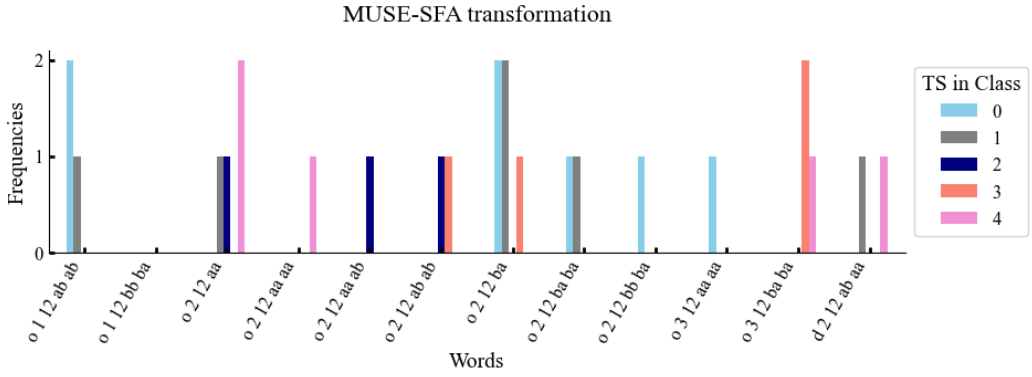

Figure 3.2B: Generated words (x axis) and their frequency (y-axis) in a five signal sample representing different category/class.

(SAX) which bins *each time series* independently, MCB bins *each time point* independently. Transformation phase of SFA applies MCB discretization such that each DFT approximation is described using those discretization's obtained from preprocessing. The result of this process is called SFA words of multivariate time series [9]. Figure 3.2A depicts this procedure of multivariate time series processing and transformation into a word representation. Similar symbolic representation methods have been developed in several studies [10].

Practical word extraction example in the figure 3.2B shows how the MUSE algorithm transforms multivariate time series of real numbers into a sequence of frequencies of words, and illustrates the features obtained for five sample time series, one of each category. For example, time series of the class 0 and 1 can be represented as a frequency vectors:

$$TS_{class=0} = [2,0,0,0,0,2,1,1,1,0,0],\ and\ TS_{class=1} = [1,0,1,0,0,0,2,1,0,0,0,1]$$

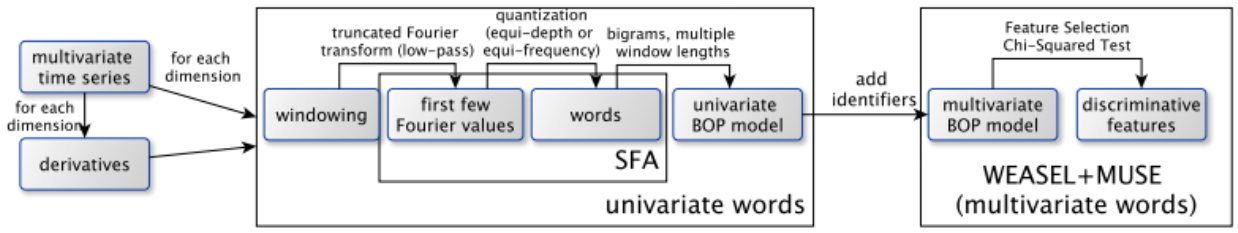

Figure. 3.3: MUSE Pipeline: Feature extraction, univariate Bag-of-Patterns (BOP) models and MUSE. [8]

Numerical class representation in the figure (TS in class) corresponds to the five dataset categories, but the actual label is not relevant in this example. Furthermore, the model setup in the word extraction example does not correspond to actual classification model setup in this study.

3.3 Classification pipeline

As a summary, in figure 3.3A the MUSE algorithm pipeline is presented as a whole. As described in the study of Schäfer and Leser, the symbolic representation SFA [9], BOP models for each dimension, feature selection and the MUSE model. MUSE conceptionally builds upon the univariate BOP model applied to each dimension. Multivariate words are obtained from the univariate words of each BOP model by concatenating each word with an identifier (representing the sensor and the window size). This maintains the association between the dimension and the feature space. [8]

3.4 Classification evaluation metrics

In analysis of actual model performance common classification evaluation metrics like overall model accuracy, precision, recall, and f1-score are used. Also, classification quality is complemented by conducting analysis of Receiver Operating Characteristic (ROC) curve and its evaluation method Area Under Curve (AUC). Further, computation time for model training and testing is considered. Model accuracy evaluation is made using *precision*, *recall*, *f1-score*, and *support* metrics.

$$Precision_c = \frac{TP}{TP + FP}, \qquad Recall_c = \frac{TP}{TP + FN}, \qquad F1score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

4. RESULTS

4.1 MUSE model optimization

MUSE algorithm hyperparameters were optimized using grid search cross validation method wherein all the possible combinations of a selected tunable parameters were tested for fitting the data. For validation a 5-fold cross validation method was used. To find optimal parameters around 40 hours were required to execute 2800 fits. The best results were gained using the hyperparameter setup depicted in table 4.1.

Table 4.1: MUSE hyperparameters

Parameter	Test values (Cross Validation)	Optimal value
alphapet size	[3,4,5,6,7,8,9]	8
window inc	[3,4,5,6,7]	5
anova	[False, True]	False
variance	[False, True]	False
bigrams	[False, True]	False
feature selection	['chi2']	chi2
p-threshold	[0.03, 0.04, 0.05, 0.06 0.07]	0.05
first order differences	[False, True]	False
support probabilities	[True]	True

Table 4.2A: Multivariate Symbolic Extension (MUSE) classification report. Used signal length is 55.

	Walking	Running	Skiing	R-Skiing	Biking	Macro Avg	Weighted Avg
Precision	0.778	0.973	0.885	0.886	0.980	0.900	0.928
Recall	0.583	0.973	0.939	0.886	0.980	0.872	0.930
F1-Score	0.668	0.973	0.911	0.886	0.980	0.883	0.928
Support	12	74	49	44	49	228	228
Accuracy	0.930						

Table 4.2B: Confusion matrix of categories for MUSE classifier. Used signal length is 55.

	Walking	Running	Biking	Skiing	Roller Skiing
Walking	58.3	16.7	8.3	8.3	8.3
Running	2.7	97.3	0	0	0
Biking	0	0	98.0	0	2.0
Skiing	0	0	0	93.9	6.1
Roller Skiing	0	0	0	11.4	88.6

4.2 Classification results

Accuracy and quality metric scores

Classification report in table 4.2A was produced using the best model setup observed in model hyperparameter optimization resulting 93,0% accuracy, however having macro average sensitivities only from 87,2% to 90,0%, especially due to a low accuracy for walking category. That can be observed also by inspecting weighted average metric which are more or less equal with accuracy.

Confusion matrix

Confusion matrix table 4.2B indicates that running and biking are quite unambiguously discriminative in the feature space. Running has been confused only with walking, although based on the data category analysis there was potential for greater confusion. Only 2,7% of running activities were classified as walking, but correspondingly 16,7% of walking activities as running.

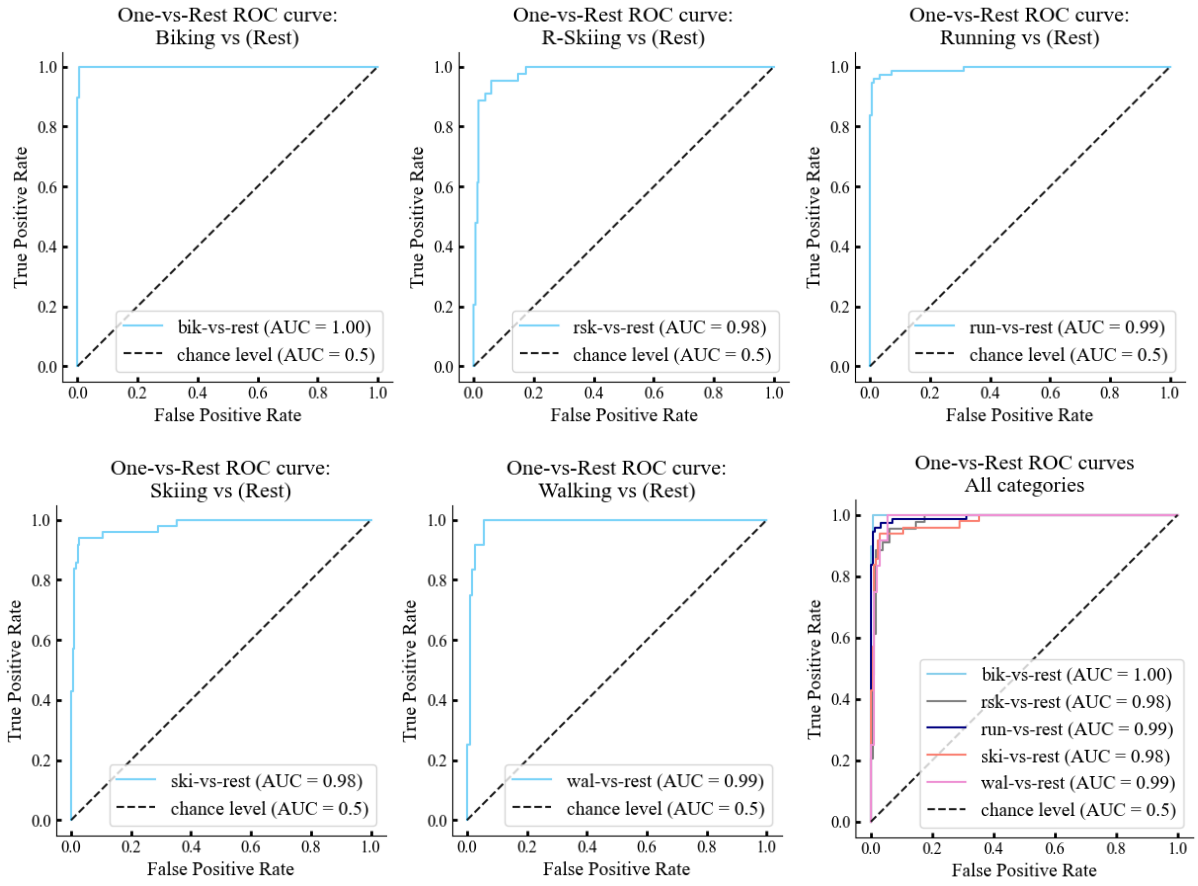

Figure 3.4: One-vs-rest ROC curves for each of the five categories. The last one combines curves into the same figure for better contrast.

Walking was all the way problematic category in overall since almost half of them were classified as something else and it was confused also with all other categories (8,3%). From the biking activities, 2,0% was classified as r-skiing which was the closest category according to pre-analysis of the data and therefore expected result. Among walking and running another challenging category pair was very similar skiing and r-skiing activities. Also, it was expected that some skiing activities might be confused with running since their heart rate and speed values may sometimes overlap. However, they were confused only together, and no single union of skiing and r-skiing activities were classified as something else.

4.3 ROC-AUC analysis

Receiver Operating Characteristics (ROC) analysis using one versus rest (1-vs-rest) method complements classification quality analysis by indicating yet more precisely which categories are the most problematic if any of them. The intent of the ROC-curve is to show how well the model works for every possible threshold, as a relation of true positive rate (TPR or sensitivity) versus false positive rate (FPR). As ROC-curve provides only a visual tool for analysis, a metric

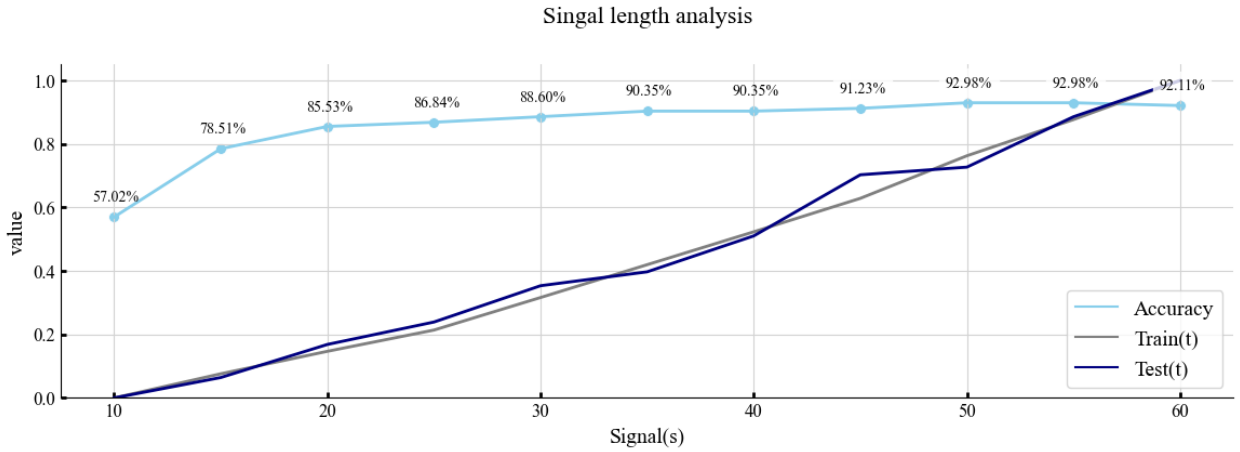

Figure 4.4: Classification accuracy and normalized training and testing times accordingly. Graph demonstrates that over 85% accuracy level is reached with only 20 seconds signals and after that point accuracy increases only moderately, whereas model computation time continues to increase almost linearly when using longer signals.

Area Under Curve (AUC) transforms it into a numerical expression allowing us to compare different ROC-curves.

In figure 3.4 ROC curves with AUC scores are shown for all the categories, complemented by a graph that combines curves into the same figure for a better observation. Also, the 1-vs-rest method naturally implies that 0.5 must be selected as a threshold value representing the worst case scenario in a binary classification and that is drawn as a dashed line. From the figure can be seen that skiing and r-skiing did not succeed as well, and their AUC score is slightly smaller (98%). That will be explained due to their mutual confusion. However, the good results for walking activity is because from the rest activities only running had 2,7% confusion with walking category while any other was predicted as walking activity. That is an important factor when we also consider the support value 12 for walking which is only $\frac{12}{216} = 5{,}6\%$.

4.4 Signal length analysis

In the signal length analysis classification was implemented using 11 different signal lengths in the range [10, 60] with a step value of 5 inspecting changes especially in accuracy and computation time accordingly. Since in the context of constructing original dataset 69 seconds signals were picked more or less intuitively conducting only a very few classification tests using an arbitrary signal length, in this study we further optimize the length of the signal and try to obtain a balance between classification accuracy and computational requirements.

Early classification

In addition to a manual classification point selection in optimal signal length evaluation as described previously we used recently developed method called Early Time Series Classification (eTSC) [11]. ETSC is the method of classifying a time series with a minimal amount of data to achieve the highest possible accuracy. The challenge of eTSC method is in tradeoff between two

Table 4.4: TEASER setup and results.

	classification points	time (s)	earliness	accuracy	harmonic mean
Train	[10, 60], step=5, 11	492	0.28	0.89	-
Test	-	-	0.33	0.86	0.76

conflicting goals, maximizing accuracy and trying to speed up classification process by determining when enough data of a time series is seen to make a decision. Using the whole available data usually improves classification accuracy but extends classification time, whereas earlier classification with less input data often leads to inferior accuracy. This kind of methods have been developed and tested in several studies during a past decade [11–13].

The TEASER algorithm [11] was applied to the data by using 11 classification points in the range [10, 60] with a step value of five. For TEASER algorithm the same optimized MUSE classifier with the same training-test data were used. TEASER consumed 492 seconds in the study test environment to determine earliness and conduct the appropriate classifications using MUSE as a slave classifier. Table 4.4 summarizes statistics regarding eTSC results. According to the results in the test data, by using 33 percent (~20/60) of the data accuracy of 86% has been achieved, which corresponds to observed accuracy in figure 4.4 in the signal point 20 seconds (85,53%). The results achieved are therefore consistent with the finding that at time point 20 the relative increase in accuracy and computation time becomes unfavorable.

5. DISCUSSION

In the discussion section we seek to analyze classification results by inspecting which kind of specific signals have been misclassified. We try to explore whether the signals are problematic in itself, possibly mislabeled in the original data, or does the applied dictionary based MUSE model, that utilizes word extraction from signals, has some fundamental weaknesses.

5.1 Misclassification analysis

In this misclassification analysis we select one interesting, misclassified activity from the predictions and try to investigate it with the help of graphics. First, mean signals of the whole dataset are generated with confidence intervals for each category and feature, as shown in figure serie 5.1A. Then the original signals for each feature have been visualized together with these mean signals. This will help us to analyze, not only in terms of single numbers, but also visually if the investigated signal has some exceptional characteristics, which may cause problems when conducting classification. In our analysis we investigate misclassified biking activity with the predicted label of R-Skiing. As can be concluded from the figure, Heart Rate feature follows unambiguously biking average, but confusion comes from Speed and Altitude dimensions. The investigated biking activity has been significantly slower than average biking activities in general. In the

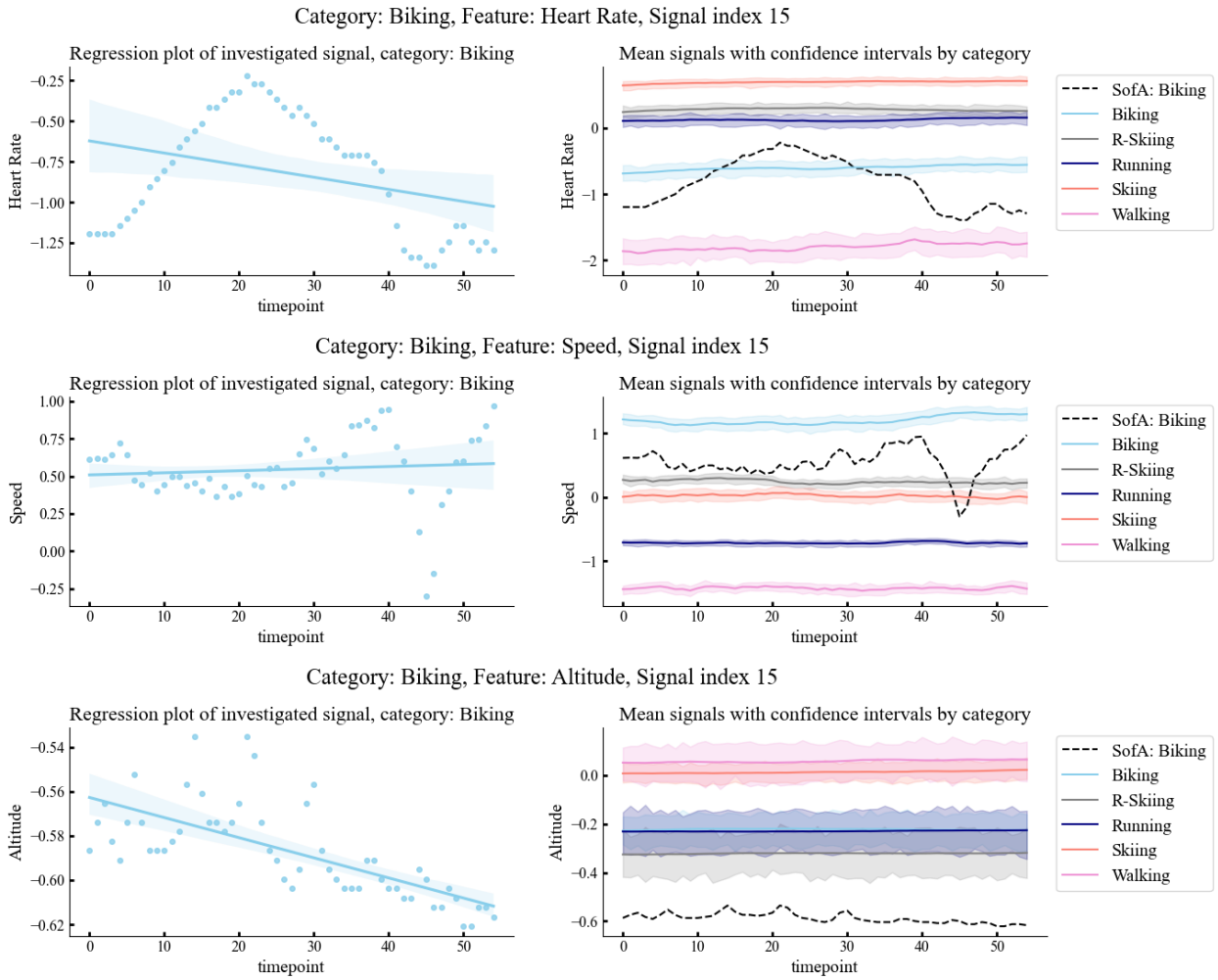

Figure 5.1A: Misclassified signal analysis. Figure serie demonstrates how different dimensions of misclassified activity signals are placed int the mean value plots with confidence intervals. Sample activity represents Biking signal which has been classified as R-Skiing. (SofA = Signal of Analysis = Misclassified signal).

same way, altitude of the biking activity has been exceptionally low, and closer to R-Skiing than the correct category.

As it might naturally happen quite often, that some biking activities may have much lower intensity implicating lower speed, these activities are poorly distinguishable using summary statistics of the signals. The situation is not better with time series statistics, where we consider interdependency of consecutive values, because that aspect neither offers sufficient discriminative characteristics between these particular activities. However, as we know that in biking activity higher speed can be maintained by a less effort implying lower heart rate, activities should be well discriminated by inspecting interdependency of features itself. In other words, when the speed value of biking activity comes closer to R-Skiing, we should determine predicted label by heart rate, or controversary. To be clear: if the heart rate value follows biking average, then if speed is considerably lower than biking average at the same time, then the activity is very likely skiing or R-Skiing. This decision making process is clearly visible in this misclassified case, Biking

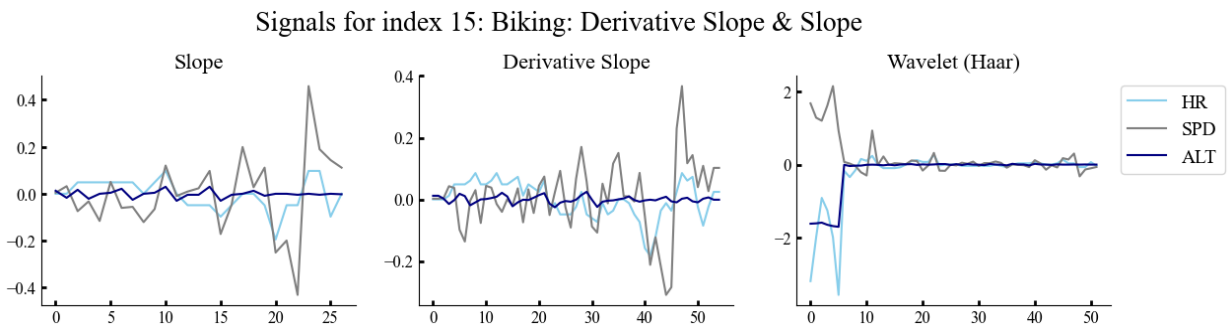

Figure 5.1B: Slope, derivative slope, and wavelet of features for investigated misclassification case: Biking predicted as R-Skiing.

predicted as R-Skiing. Therefore, the mistake cannot be taken as a model weakness, but rather problematic sport activity, that might not possibly be classified correctly using study dataset. It is also reasonable to mention a known fact, that in the applied dataset biking and roller skiing activities have been performed in very similar conditions: the same road sections with the similar altitude value, altitude derivatives etc. We may extract shapelet features, such as slope, derivative slope, or wavelet as in figure 5.1B, but we cannot expect remarkable improvement on the basis of well-known background facts.

5.2 Discussion summary

It could be suggested that heart rate responds to uphill's in quite different manner among sports providing clearly discriminative factor to classify these specific outdoor sport activities. Reflecting thoughts to investigated Biking predicted as R-Skiing error, probably the most potential way to distinguish these activities is to investigate that heart rate response when altitude is decreasing or increasing. If the heart rate time serie sensitively follows changes in altitude time serie, activity is highly likely R-Skiing regardless of the athlete. However, according to the results of the study [add here], MUSE has been the most successful model – that was also the main reason to select it for more thorough study – which indicates that shapelet or wavelet based models do not work as effectively, at least, in the univariate time series of the same data. Also, the dataset in this study was transformed to differential values without success when using MUSE. Therefore, we should develop other than dictionary based multivariate model for this purpose, which considers interdependency of derivative values of the features.

Complicating classification model without domain knowledge may not provide desired results. By developing an accurate model using the knowledge about, for example, that the heart rate responses to uphill's distinctively among sports, we could develop better functioning model. Despite MUSE constructs words of the multivariate signals considering interdependency of the dimensions, it does not seem to provide sufficient built-in method to consider, for example, derivatives of the features and their correlations, and thus separate data transformations are needed. In the investigated misclassification incident sufficient information in the data quite obviously exist but currently MUSE is not able to extract that information from the signal in order to help to

improve accuracy of the classification. That might also reveal the possible fact that segment selection from the original data becomes extremely crucial when using the dataset collected in uncontrolled conditions. However, the original idea was to classify sport using time series data from the premise where we have knowledge about the diverging correlation of these features, but in order to further improve classification results, it might require clear dataset of sport activities recorded in controlled environment. Therefore, the next step could be to collect similar dataset which can be used to optimize algorithm development for this purpose, and then compare achieved results, and further to apply trained model into the dataset used in this study.

6. CONCLUSIONS

The results of the study indicate that multivariate time series classification can provide a well performing method to classify a newly introduced multidimensional time series dataset of sport activities, recorded by an individual athlete. The MUSE, which utilizes symbolic representation of signals using SFA transformation, was validated to be well applicable. When using genuine data from uncontrolled environment it produced 93% accuracy with 50 seconds signals. Further the results of the study show that the introduced dataset does not seem to have significant amount of problematic individual instances which may interfere the classification but many of the misclassified signals could be classified correctly using traditional CML with extracted summary features from the signals.

Future research may employ a variety of alternative algorithms, including CML and well-known neural networks, on the same data in order to contrast the findings. The dataset might also be used in any TSC research to assess model performance, not just in terms of the data from the controlled experiment but also in terms of how well the algorithms function when applied to the newly introduced dataset from the intended real-world application. On the other hand, introduced dataset undoubtedly offers an opportunity to develop data pre-processing methods with better outcomes through more meticulous data cleansing before actual classification.

7. REFERENCES

1. Demrozi F. et al. Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey // IEEE Access. 2020. Vol. 8. P. 210816–210836.
2. Lara O.D., Labrador M.A. A Survey on Human Activity Recognition using Wearable Sensors // IEEE Commun. Surv. Tutor. 2013. Vol. 15, № 3. P. 1192–1209.
3. Berchtold M. et al. An Extensible Modular Recognition Concept That Makes Activity Recognition Practical // KI 2010: Advances in Artificial Intelligence / ed. Dillmann R. et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. Vol. 6359. P. 400–409.
4. Lara Ó.D. et al. Centinela: A human activity recognition system based on acceleration and vital sign data // Pervasive Mob. Comput. 2012. Vol. 8, № 5. P. 717–729.
5. Tapia E.M. et al. Real-Time Recognition of Physical Activities and Their Intensities Using Wireless Accelerometers and a Heart Rate Monitor // 2007 11th IEEE International Symposium on Wearable Computers. Boston, MA, USA: IEEE, 2007. P. 1–4.

6. Foerster F., Smeja M., Fahrenberg J. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring // Comput. Hum. Behav. 1999. Vol. 15, № 5. P. 571–583.
7. Löning M. et al. Sktime: A Unified Interface for Machine Learning with Time Series. arXiv, 2019.
8. Schäfer P., Leser U. Multivariate Time Series Classification with WEASEL+MUSE. arXiv, 2017.
9. Schäfer P., Högqvist M. SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets // Proceedings of the 15th International Conference on Extending Database Technology. Berlin Germany: ACM, 2012. P. 516–527.
10. Baydogan M.G., Runger G. Learning a symbolic representation for multivariate time series classification // Data Min. Knowl. Discov. 2015. Vol. 29, № 2. P. 400–422.
11. Schäfer P., Leser U. TEASER: early and accurate time series classification // Data Min. Knowl. Discov. 2020. Vol. 34, № 5. P. 1336–1362.
12. Xing Z., Pei J., Yu P.S. Early classification on time series // Knowl. Inf. Syst. 2012. Vol. 31, № 1. P. 105–127.
13. Tavenard R., Malinowski S. Cost-Aware Early Classification of Time Series // Machine Learning and Knowledge Discovery in Databases / ed. Frasconi P. et al. Cham: Springer International Publishing, 2016. Vol. 9851. P. 632–647.