Comparator Model for Detecting Changes in the Ease of Breathing of COPD Patients

Thomas T. Kok[1,*], Willemijn Groenendaal[2], Dolores Blanco-Almazán[3,4,5], Lien Lijnen[6], Christophe Smeets[7], David Ruttens[6,8], John Morales[2], Tom Dhaene[1], Femke Ongenae[1], Sofie Van Hoecke[1] and Dirk Deschrijver[1]

[1]IDLab, Ghent University - imec

[2]Imec The Netherlands/Holst Centre

[3]Department of Automatic Control (ESAII) at the Universitat Politècnica de Catalunya (UPC)-BarcelonaTech

[4]Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology

[5]Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN)

[6]Department of Medicine and Life Sciences, Hasselt University

[7]Future Health department, Ziekenhuis Oost-Limburg

[8]Department of Respiratory Medicine, Ziekenhuis Oost-Limburg

Abstract

This paper introduces a new machine learning based comparator model to assess changes in the ease of breathing of COPD patients during loaded breathing. The comparator model is based on a random forest classifier that detects whether breathing becomes either more difficult, easier or remains stable. The designed model can accurately detect respiratory changes by comparing temporal segments of physiological signals measured during loaded breathing, with an F_1 score of almost 80%, resp. 70% for the wearable solution. As the model is trained and tested with features derived from different signal modalities, such as respiratory flow, audio, bio-impedance and accelerometer data, we also did a systematic comparison of the signal modalities to assess their predictive power.

Keywords

machine learning, biomedical signals, respiratory status, COPD, ease of breathing prediction

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease where the airflow from the lungs is obstructed. This can lead to several adverse symptoms, such as breathing difficulty and shortness of breath, excess phlegm or sputum, and frequent coughing or wheezing. In the United States, it is estimated that 16 million Americans have this disease, whereas millions more people suffer from COPD but have not yet been diagnosed [1]. As a result, morbidity and mortality in COPD patients are considerably high. In order to diagnose COPD, a spirometer test is applied as a gold standard test for pulmonary function. The patient inhales and then exhales with a maximal effort through a mouthpiece, while the airflow going into and out of the lungs is measured and analyzed [2].

A limitation of this test is that it has to be performed by trained medical personnel, and thus does not allow longitudinal monitoring of the respiratory status, which is desirable, because early detection and treatment of exacerbations can prevent disease worsening [3].

As such, there is a need for the development of a wearable respiration monitoring system that uses information captured in a real-world setting, i.e. outside of spirometry, to allow new steps in automatically and continuously analyzing the patient's status. While diagnosing COPD severity is out of scope for such a potential system, longitudinally monitoring changes in respiratory status or ease of breathing could be used as an additional tool in monitoring and treatment of COPD.

Several approaches have been devised to assess the patient's condition, by measuring information such as heart rate, oxygen saturation and respiratory flow [4]. Some of these approaches are based on analyzing the respiratory rate, rather than taking into account the entire respiratory pattern [5, 6]. It has also been shown that significant correlations are present between non-invasive physiological signals that can be recorded with wearables, such as thoracic bioimpedance, electrocardiogram, or photoplethysmogram, and parameters that are relevant to assess the respiratory status of a patient, such as the respiratory volume or respiratory muscle disfunction [7]. Hence, signals recorded with wearable devices may contain relevant information to detect respiratory changes in COPD patients. A systematic comparison of the predictive power of different physiological signals is
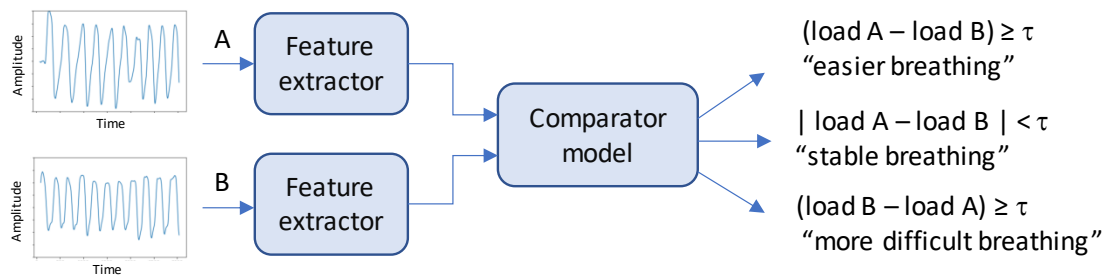
✉ thomas.kok@ugent.be (T. T. Kok)

Figure 1: Outline of comparator model that discriminates between an increase, stability and decrease in ease of breathing, due to changes in inspiratory loading.

still largely unaddressed.

This paper aims to address and explore this by proposing a new machine learning based comparator model that detects whether breathing becomes either more difficult, easier or remains stable. The patient's ease of breathing is assessed by comparing pairs of segments of non-invasive physiological signals that were recorded during a loaded breathing test. By analysing the performance of the model, valuable new insights are obtained on the choice of signals that are most relevant to monitor changes in respiration. As such, the outcome of this study provides useful insights to decide which signal modalities should be measured by the wearable device, and forms a basis for the development of next-generation wearable respiratory monitoring technology. The availability of such a technology can possibly result in a faster intervention to prevent disease worsening, potentially leading to reduced health care costs, hospital (re)admissions and improved quality of life.

In summary, the main contributions of this paper are threefold:

- A systematic analysis and comparative study of multiple physiological signals is performed, to identify which ones are most influential to correctly detect changes in the ease of breathing of a COPD patient during inspiratory loading.
- Relevant features are automatically extracted from the signals, and a state-of-the-art classification algorithm for time series data is used to build a comparator model that accurately detects whether breathing becomes either more difficult, easier or remains stable.
- The performance of the model is validated on data from a clinical study and the results are evaluated.

2. Materials and Methods

2.1. Data collection and setup

The study uses data gathered from 50 patients with COPD and enrolled for an inspiratory loaded breathing test in a clinical setting at Ziekenhuis Oost-Limburg (Belgium). The data from six patients were excluded in this study due to one or more of the collected signal modalities being either unavailable or saturated. The resulting data set comprises 44 patients.

The study was approved by the local institutional medical ethics committee from Ziekenhuis Oost-Limburg with reference 18/0047U. The study followed the World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Humans Subjects. All patients provided written informed consent.

An incremental inspiratory threshold loading protocol was performed, where the patients were imposed to increasing inspiratory loads proportional to their maximal inspiratory pressure (MIP). The loads are quantified as 0%, 12%, 24%, 36%, 48% and 60% of the MIP that was measured before the start of the test [8]. As such, a higher load corresponds to an overall higher inspiratory effort that is required to breathe. Each load test was applied for a period of thirty breaths (i.e. a variable time length), with a two-minute resting period between the tests for each different load. The loads were sequentially applied in increasing order. The measurement of the MIP and the imposition of the loads was both done using an inspiratory muscle trainer (POWERbreathe KH2, POWERbreathe International Ltd, Southam, UK).

During the test, two systems were used to record several physiological signals simultaneously. The first system was a standard wired acquisition system (MP150, Biopac Systems, Inc., Goleta, CA, USA), used to record the accelerometer data and audio signals; the airflow was measured using Biopac, together with a pneumotach transducer (TSD107B, Biopac Systems, Inc.) connected to a differential amplifier (DA100C, Biopac Systems, Inc.). Given the fact that the lung sound signals have a band-

width around 4000 Hz [9], the Biopac was configured to record at a sampling frequency of 10,000 Hz, and this was applied to all the channels. The second system was a low-power wearable device (imec the Netherlands, Eindhoven, the Netherlands) with an injecting current of 100 uAp-p at 80 kHz that was used to record the bio-impedance signal with a sampling frequency of 16 Hz [10]. The location and configuration of the signals were as follows:

- Respiratory flow (gold-standard, used as a baseline)
- Accelerometer: parasternal and diaphragm (lower intercostal spaces), three axes
- Microphone (audio): left lung, right lung, tracheal
- Bio-impedance (bioZ): 4 configurations

For the bio-impedance signals, the first of the four channels described in [10] is used in this study, because it was shown to have a more robust performance for estimating respiratory volume changes when higher loads are imposed. The full details about this data collection protocol for the respiratory flow, bio-impedance, and accelerometer signals, as well as the set-up of the experiments are explained in [7, 11]. Regarding the audio signals, these were recorded using three microphones (TSD108, Biopac Systems, Inc) with a frequency response of 35-3500 Hz. Two microphones (for both lungs) were positioned on the back, two to three centimeters below the shoulder blades, at each side of the spinal cord. The other microphone (for the tracheal sound) was positioned on the right side of the patient's neck. After amplifying the sound 200 times, it was filtered with an analog low-pass filter of 5 kHz and a high-pass filter of 0.05 Hz. In the forthcoming sections of the paper, the respiratory flow is referred to as an obtrusive signal modality, whereas all others are considered as unobtrusive signal modalities.

2.2. Preprocessing and segmentation

The signals contain noise due to subject movement, electrical inference, measurement noise and other disturbances. In order to extract all relevant information, all signals except audio are first filtered with a low-pass zero-phase Butterworth filter with the following orders and cut off frequencies per signal modality: a fifth-order 40 Hz low-pass filter for the respiratory flow signal, a fourth-order 0.7 Hz low-pass filter for the bio-impedance signal, and a eighth-order 40 Hz low-pass filter for the accelerometer data. The filter values are based on the characteristics of the noise present in the signals, removing higher frequency noise while retaining the relevant information below the cut off frequencies.

All signals are resampled to a common sampling rate. The respiratory flow, accelerometer, and audio signals are downsampled from 10,000 Hz to 100 Hz, whereas the bio-impedance signals are upsampled using cubic interpolation from 16 Hz to 100 Hz.

After filtering, a bidirectional moving average and moving variance filter is applied to remove artefacts in the data where the patient did not breathe into the pneumotach transducer correctly. When either of these values in either direction is close to zero, this is considered an artefact and removed.

After removal of artefacts, the signal is divided into multiple shorter signals. In order to define a common-length input signal for the machine learning model, each part of the signal that is not corrupted, is subdivided into several smaller segments having a predefined window length w. For each combination of patient and inspiratory load, one window is sampled from the longest stable breathing period without interruptions or artefacts. This ensures that each segment has a consistent length.

It was found that a window length of $w = 30$ sec is the most adequate choice, because it ensures that at least 95% of the input space (i.e. all possible patient-load combinations) is still included in the dataset after preprocessing, while guaranteeing the inclusion of multiple breaths. This window length strikes the balance between including as much data as possible in the entire window and including as many patient-load combinations as possible, avoiding biased results.

2.3. Machine learning-based comparator model

Fig. 1 shows an overview of the designed comparator model. Pairs of segments from the same signal modalities of the same patient are used as an input for the model, and a three-way classification is calculated, depending on whether the load of the second signal is lower than, higher than, or comparable to the load of the first signal. It is hypothesized that this method can then also be used to detect increased difficulty in breathing, which may be indicative for worsening of the respiratory condition. The threshold τ used to separate the categories is set to 12%, to match the granularity of variations in the loads.

For each signal type, a comparator model is trained using pairs of signals recorded at different load combinations, with the labels corresponding to an increase, decrease, or stability in inspiratory load. Since the model is trained on data from all the patients in the training set at once, a general model is obtained, rather than a patient-specific one.

As six different loads were considered in the breathing test, 21 pairs of load combinations can be generated. The swapped pairs are also included to train the model, resulting in 42 load pairs. As such, it would have been possible to have 1848 pairs. However, one patient was able to perform only 5 loads instead of 6. For this reason, 1836 pairs were available for the development of the model.

2.3.1. Feature extraction

The tsfresh package [12] is used to compute a set of 783 features f_i from the time series signals, both in the time and frequency domains, generated from 74 unique features with varying parameters (e.g., the window size, the quantiles, etc.). They include statistical features or combinations thereof, and more complex features. A comprehensive overview is provided in Appendix A.

The features are then filtered, keeping only those features that are determined as relevant for predicting the label within the training data [12].

In the case of accelerometer data, the signal modality is multivariate (tri-axial). As such, features were extracted and filtered from the data along each coordinate axis separately. All other signal modalities are considered as univariate.

The difference between the normalized values of the same features $f_i(A) - f_i(B)$, is computed for every combination of two loads, belonging to the same patient, and used as training input for a Random Forest classifier [13]. This model type is known to provide good results in other medical studies [14], and its performance is relatively stable due to the insensitivity to noise or overfitting, allowing a fair comparison between signals. For the same reason, and with the goal of comparing signal modalities rather than finding the absolute best model, generic parameters are used: the classifier consists of 100 decision trees, with no restrictions on the depth, number of leaf nodes or splits. For splits, the Gini impurity is used along with a maximum of $\sqrt{n}$ features considered for each split.

By explicitly considering the difference, rather than providing the union, a more compact representation is obtained, and the model is enforced to learn that the same characteristics between both signals should be compared. By running experiments for different configurations of feature extraction method and signal modalities, it was experimentally verified that this approach leads to an F_1 score that is consistently higher.

2.3.2. Model training and cross-validation

The performance of the models is evaluated using 10-fold cross-validation, ensuring no overlap of samples from patients between different folds. For each training and test step, features are extracted from the 9 training folds with the exact same parameters each time. The model performance is assessed for every pair of loads from patients in the test fold. Note that the evaluation is based on a weighted F_1 score to take into account the small class imbalance, as instances from the stable condition are less common than from the worsening and improving conditions.
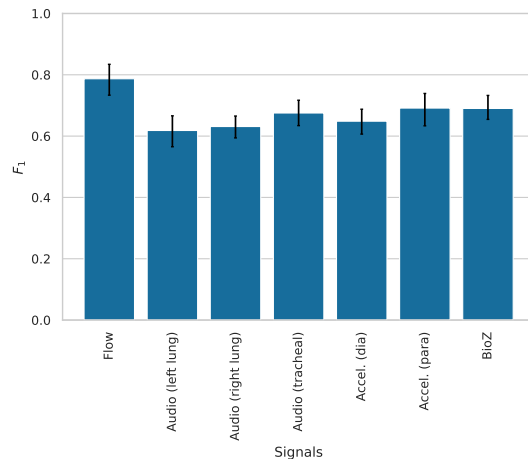

Figure 2: Weighted F_1 score (with 95% confidence interval) for different input signal modalities and comparator models
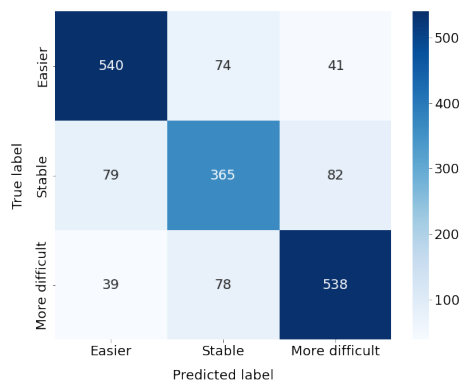

Figure 3: Confusion matrix of the best performing model, where each row/column represents the true/predicted instances for a class. The class labels refer to changes in ease of breathing due to inspiratory loading. The majority of the instances are classified correctly. Misclassifications are more common between neighbouring classes.

3. Results

An overview of the model performance for various signal modalities is shown in Fig. 2.

A one-way ANOVA analysis is performed on all results of the unobtrusive signal modalities, with $\alpha = 0.05$ [15]. For this, all F_1 scores are considered from each signal modality as a separate group, with the null hypothesis that the means for each group are sampled from the same distribution. The resulting p-value of $0.029 < \alpha$ and rejection of this null hypothesis confirms that there is a

Class	Sens.	Spec.	PPV	NPV	Acc.	F_1
Easier	0.82	0.90	0.82	0.90	0.87	0.82
Stable	0.69	0.88	0.71	0.88	0.83	0.70
More difficult	0.82	0.90	0.81	0.90	0.87	0.82

Table 1
Quality metrics of best performing model (sensitivity, specificity, positive predictive value, negative predictive value, accuracy and F_1 score)

statistically significant difference in the mean F_1 scores of the signal modalities. Amongst those signal modalities, bioZ, accelerometer (parasternal) and audio (tracheal) seem to provide the best predictive power to detect changes in the ease of breathing. A paired t-test with $\alpha = 0.05$ confirms that the F_1 score for the tracheal audio signal is significantly higher than the left lung ($p = 0.002$) and right lung ($p = 0.028$) signals. Furthermore, the F_1 score for parasternal accelerometer signal is significantly higher than the diaphragm accelerometer signal ($p = 0.025$). There is no significant difference between the F_1 score of the bioZ and accelerometer (parasternal) signals ($p = 0.821$), the bioZ and audio (tracheal) signals ($p = 0.401$), and the accelerometer (parasternal) and audio (tracheal) signals ($p = 0.64$).

Fig. 3 shows the confusion matrix of the overall best performing model, i.e. the random forest model based on tsfresh features, that considers the (obtrusive) respiratory flow signal. It is seen that the majority of the instances are classified correctly, and misclassifications are more common between neighbouring classes. Table 1 provides an overview of the performance metrics calculated to assess the quality of the model, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and the F_1 metric for each of the three classes. From both Fig. 3 and Table 1, it is seen that changes in breathing difficulty can be identified more accurately than stability. This is not unexpected, as the stable class is more similar to the other two classes, than they are to each other.

4. Discussion

The results confirm that differences in the respiratory pattern during the inspiratory load protocol applied to COPD patients can be assessed by our machine learning based comparator model. The best performance is obtained when considering the respiratory flow signal, with an F_1 score of 0.78. This demonstrates the ability of the model to discriminate between an increase, remaining stable or decrease in the ease of breathing for patients.

However, there is still an error margin associated with the classification performance of the model. For example, the model has a 10% chance of missing an increase in difficulty and a 19% chance of being wrong when an increase is detected. As such, it is important to interpret the classification results carefully and weigh the potential consequences of false positives and false negatives, as the acceptable margin of error may vary depending on its application. For example, this error margin may be acceptable for regular monitoring of stable COPD patients, while it may not be acceptable for high-risk patients.

When considering other signals that can easily be acquired with wearable devices, a lower F_1-score is observed when compared to the use of the respiratory flow signal. Nevertheless, the actual performance of these models is less important from a clinical perspective, because the comparator model in this study is only used to benchmark and rank the different signal modalities according to their predictive power. Having these insights can be valuable, because an identification of the best-performing signal modalities can help to make an informed choice of sensors during the design of a wearable.

Such a wearable can collect longitudinal data from patients during normal daily activities, on which the comparator model can be retrained. Having more lengthy signals makes it possible to consider multiple window segments, which can further boost the model performance. Furthermore, the availability of more extensive data sets creates new possibilities to apply advanced deep learning techniques that have shown to be effective on similar problem settings with respiratory data [16], while also enhancing the generalizability of the current feature-based model.

Future work will focus on the identification of an optimal combination of unobtrusive signal modalities to avoid redundancy in the selection of signal modalities within a certain category. Longitudinal, clinical and external validation of the approach will also be performed. Additionally, to increase trust in the model, interpretable machine learning methods will be explored. Generating explanations alongside classifications also allows for physicians to incorporate the reasoning of the model within their own decisions.

5. Conclusion

This paper presents a novel machine-learning based comparator model that detects changes in the ease of breathing of COPD patients during inspiratory loaded breathing. Numerical results provide a comparison of different input signals and models. When applied to the respiratory flow, a weighted F_1 score of 0.78 is obtained. When considering other signal modalities that are not as obtrusive, and can be measured with wearable devices, the ones that offer

the best predictive performance are bioZ, accelerometer (parasternal) and audio (tracheal), with a weighted F_1 score of 0.69, 0.69 and 0.68 respectively.

Acknowledgments

References

[1] N. Tian, M. Zack, A. Wheaton, K. Greenlund, J. Croft, Epilepsy and chronic obstructive pulmonary disease among u.s. adults: National health interview survey 2013, 2015, and 2017., Epilepsy Behav. 110 (2020) 107175.

[2] B. L. Graham, I. Steenbruggen, M. R. Miller, I. Z. Barjaktarevic, B. G. Cooper, G. L. Hall, T. S. Hallstrand, D. A. Kaminsky, K. McCarthy, M. C. McCormack, C. E. Oropez, M. Rosenfeld, S. Stanojevic, M. P. Swanney, B. R. Thompson, Standardization of spirometry 2019 update. an official american thoracic society and european respiratory society technical statement, American Journal of Respiratory and Critical Care Medicine 200 (2019) e70–e88.

[3] J. Vestbo, S. Hurd, A. Agusti, P. Jones, C. Vogelmeier, A. Anzueto, P. Barnes, L. Fabbri, F. Martinez, M. Nishimure, R. Stockley, D. Sin, R. Rodriguez-Roisin, Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease: Gold executive summary, Am. J. Respir. Crit. Care Med. 187 (2013) 347–365.

[4] A. Rajeh, J. Hurst, Monitoring of physiological parameters to predict exacerbations of chronic obstructive pulmonary disease (copd): A systematic review, Journal of Clinical Medicine 5 (2016) 108. doi:10.3390/jcm5120108.

[5] A. Yanez, D. Guerrero, R. Perez de Alejo, F. Garcia-Rio, J. Alvarez-Sala, C.-R. M., R. Malo de Molina, M. Valle Falcones, P. Ussetti, J. Sauleda, E. Zamora Garcia, J. Rodríguez-Gonzalez-Moro, M. Franco Gay, M. Torrent, A. Agusti, Monitoring breathing rate at home allows early identification of copd exacerbations, Chest 142 (2012) 1524–1529.

[6] J.-C. Borel, J. Pelletier, N. Taleux, A. Briault, N. Arnol, C. Pison, R. Tamisier, J.-F. Timsit, J.-L. Pepin, Parameters recorded by software of non-invasive ventilators predict copd exacerbation: a proof-of-concept study, Thorax 70 (2015) 284–285.

[7] D. Blanco-Almazán, W. Groenendaal, M. Lozana Garcia, L. Lijnen, C. Smeets, D. Ruttens, F. Catthoor, R. Jané, Combining bioimpedance and myographic signals for the assessment of copd during loaded breathing, IEEE Trans. on Biomedical Engineering 68 (2020) 298–307.

[8] ATS/ERS, Ats/ers statement on respiratory muscle testing, American Journal of Respiratory and Critical Care Medicine 166 (2002) 518–624.

[9] S. Reichert, R. Gass, C. Brandt, E. Andrès, Analysis of respiratory sounds: state of the art, Clin Med Circ Respirat Pulm Med 2 (2008) 45–58.

[10] T. V. Steenkiste, W. Groenendaal, P. Dreesen, S. Lee, S. Klerckx, R. de Francisco, D. Deschrijver, T. Dhaene, Portable detection of apnea and hypopnea events using bio-impedance of the chest and deep learning, IEEE Journal of Biomedical and Health Informatics 24 (2020) 2589–2598.

[11] D. Blanco-Almazán, W. Groenendaal, F. Catthoor, R. Jané, Wearable bioimpedance measurement for respiratory monitoring during inspiratory loading, IEEE Access 7 (2019) 89487–89496.

[12] M. Christ, N. Braun, J. Neuffer, A. Kempa-Liehr, Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package), Neurocomputing 307 (2018).

[13] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[14] N. Emanet, H. Oz, N. Bayram, D. Delen, A comparative analysis of machine learning methods for classification type decision problems in healthcare, Decision Analytics (2014) 1–20.

[15] J. McDonald, Handbook of Biological Statistics, 3 ed., Sparky House Publishing, Baltimore, Maryland, 2014.

[16] O. Pastor-Serrano, D. Lathouwers, Z. Perkó, A semi-supervised autoencoder framework for joint generation and classification of breathing, Computer Methods and Programs in Biomedicine 209 (2021) 106312.

A. Overview of extracted features

Simple features	
mean (μ)	median
standard deviation (σ)	variance (σ^2)
(absolute) minimum	(absolute) maximum
absolute energy	root mean square
quantiles (10th)	skewness
kurtosis	value and range count

Combinations of simple features	
$\sigma^2 > \sigma$	$\sigma > r \cdot (\text{max - min})$
standard error $\cdot \frac{1}{\mu}$	duplicate value (max / min)
average over differences	mean of second derivative
longest subsequence above and below mean	count above and below mean
sum of reoccuring values	percentage of reoccuring values
symmetry	percentage of unique values
index of of mass quantiles (10th)	number of zero-crossings
energy ratio by chunks	$r \cdot \sigma > \mu$
mean of n largest values	first and last location of minimum
first and last location of maximum	

Complex features	
Benford correlation	time reversal asymmetry statistic
c3 statistic	complexity-invariant distance
autocorrelation statistics	number of (unique) peaks
binned entropy	permutation entropy
fourier entropy	CWT coefficients
cross power spectral density	autoregressive process coefficient
FFT statistics	Langevin coefficients
linear trend statistics	augmented Dickey-Fuller test
Lempel-Ziv complexity estimate	matrix profile