A New Frequency Domain and Dynamic Time Warping Based Feature: WFOD Feature

Ekin Can Erkus[1, a)] and Vilda Purutcuoglu[2, b)]

[1)]*Department of Biomedical Engineering, Middle East Technical University, Ankara, Turkey*
[2)]*Department of Statistics, Department of Biomedical Engineering, Middle East Technical University, Ankara, Turkey*

[a)]*Corresponding author: eerkus@metu.edu.tr*
[b)]*Electronic mail: vpurutcu@metu.edu.tr*

Abstract. Anomaly detection is one of the critical steps in the diagnosis of disorders. Many anomaly detection methods aim to use different characteristics of the anomalous data with distinctive features than the normal data. We propose a new feature extraction method, the WFOD feature which is based on a novel data transformation in the frequency domain called WFOD. In our analyses, we use a publicly available motion artifact ECG dataset, which are collected in three different cases: standing, walking and jumping. Such cases are classified by four different classifiers with the pairs of statistical moments of the data, with and without the WFOD feature. The results show that the WFOD feature enhances the classification accuracy in most cases by improving accuracy by up to 25% on accuracy values ranging from the worst case, 47% to the best, 93%

INTRODUCTION

Anomalies are defined as the components or intervals in data which behave irregularly compared to the rest of the data [1]. Therefore, anomaly detection is one of the main steps in detailed data analyses in distinct applications from classification via machine learning approaches, or diagnosis in biomedical applications [2]. On the other hand, the detection of the anomalies is often highly dependent on the data features, which make the selection and extraction of the features from the data, another important step. Most data modalities have specific types of distinguishing features, but often basic statistical features such as the statistical moments are also used as features in the classification studies [3].

Apart from the time domain features, frequency domain features are also useful and widely used in many engineering applications to discriminate data [4]. Time domain features are highly dependent on data trends and environmental artifacts such as motion, whereas frequency domain features are not as much affected by trends or short time anomalies [5]. On the other hand, frequency domain features are highly susceptible to periodic noises such as hum noise or periodic muscle contractions, while time domain features do not change much with such noises with low signal-to-noise ratio [6]. Therefore, hybrid features representing both time and frequency domain properties can provide a better broad sense of applicability on time series data modalities. However, since they include a transformation and several steps to compute, such hybrid features are generally computationally complex, making them not completely useful for real-time analyses.

In this study, a new hybrid feature with a novel data transformation, the WFOD feature which has both time and frequency domain properties, is proposed. The performance of the new feature is tested by using an open-access ECG motion artifact dataset in a classification study. The statistical moments, namely, mean, variance, skewness, and kurtosis, are paired with each other in combination with and without the WFOD feature. Moreover, four classifiers, namely, tree classifier, linear discrimination analysis (LDA), K-nearest neighbor (K-NN), and Naive Bayes classifier, classify the features independently. The results are compared with each other, and the performance of the WFOD feature in terms of discrimination of the ECG data of different cases is investigated.

PROPOSED FEATURE: WFOD TRANSFORM AND FEATURE EXTRACTION

WFOD algorithm consists of several steps in both the time and frequency domain, and based on the frequency domain based outlier detection method (FOD) [7]. Firstly, the periodogram of the data with defined window size, is estimated. Such window size, is initially selected to cover at least five heartbeats to generate clear patterns in the frequency domain. The second step of the algorithm is to find the principal frequency in the periodogram plot. Both to detect the principal frequency and to discard the extreme or possibly erroneous cases in the data, such as the heartbeat above 240bpm and below 30bpm, the searching window in periodogram plot is limited between $f_{int}max$ and $f_{int}min$,

respectively. The formula to compute $f_{int}max$ can be found in Equation 1, where the constant 0.25 corresponds to the time between the beats of 240bpm. Similarly, Equation 2 represent the formula to calculate $f_{int}min$ value, where the constant 2 stands for the time interval between the beats of 30bpm.

$$f_{int}max = \frac{DataLength}{0.25 * SamplingRate} \tag{1}$$

$$f_{int}min = \frac{DataLength}{2 * SamplingRate} \tag{2}$$

Finding the principal frequency may be tricky for some data with multiple periodic behavior or noise. Hence, the next step is to detect the peaks between $f_{int}min$ and $f_{int}max$ to provide a somewhat selective approach to detect the correct peak in the defined frequency interval. Here, the peak with the highest amplitude and the first peak are selected. If they correspond to the same frequency, then that frequency is decided to be the principal frequency, f_{int}. If not, the frequency value with the highest peak amplitude is decided. A note for future work should be here to improve the detection of the principal frequency. Hence, by using f_{int} and the window size in the time domain, w, the common time domain periodicity, tint can be found by the application of the relationship between the time domain and the frequency domain by using Equation 3.

$$t_{int} = \frac{2 * W}{f_{int}} \tag{3}$$

Hence, if the data in the window are regular and do not include any secondary oscillation or artifact, the peak can easily be detected. On the other hand, if the data have an anomaly or artifact, then the location of the peak changes or, the peak cannot be found. Such cases indicate anomalies in the current time window in which the algorithm is being iterated. Therefore, the values of tint can be a pre-indicator of the various types of anomalies. Moreover, t_{int} value can also be used in the transformation of the data into a more basic and stationary version, which we call as 'WFOD transform'. The transformed data consist of the zero vector with length equal to the window size and single-point impulse peaks with amplitude equal to the global extreme value of either minimum or maximum, whichever has the highest absolute amplitude value, of the data in the window. Such single-point impulse peaks are located equidistant of each other by the interval of t_{int}. Hence, the data window can be represented simply by the common periodicity and the impulses. The WFOD transformation algorithm uses the following steps:

1. Define the parameter of the window size.
2. Obtain t_{int} from Equation 3.
3. Build a zero vector whose length matches the length of window size.
4. Obtain the global extremum value, i.e., $GE_{amplitude}$, and its location, i.e., $GE_{location}$, in current window.
5. Add impulses with the amplitude $GE_{amplitude}$ at the location $GE_{location}$ on zero vector.
6. Shift the location backwards and forwards by t_{int} on the zero vector, and place additional impulses.
7. WFOD transformation is obtained.

A representation of the application of WFOD transformation on real data [8] can be found in Figure 1, where the data are represented with the blue line, and their WFOD transformation is represented with the red line. The equidistant impulses with the distance of t_{int} and the amplitude of $GE_{amplitude}$ can be observed in this figure too.

After the computation of WFOD transformation, to extract features from this transformation, a dynamic time warping algorithm (DTW) between the original data and the WFOD transformed data is applied. Hence, the DTW value is implemented as a new feature, representing both the data's time and frequency domain characteristics.

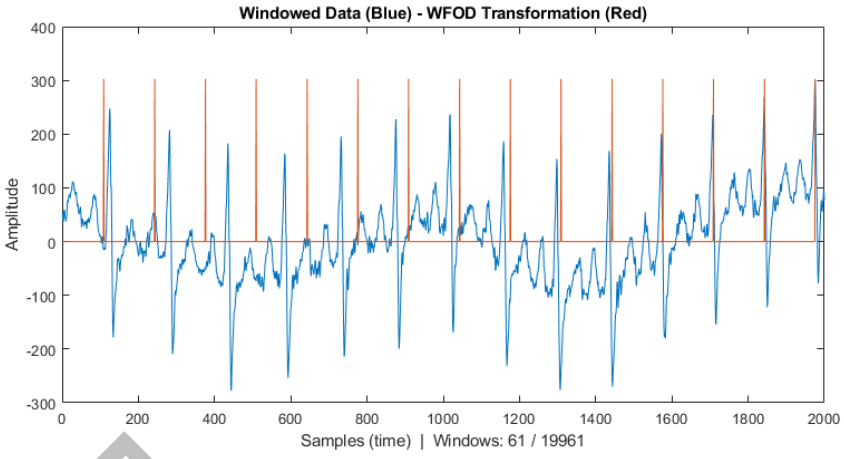

FIGURE 1. A representation of the windowed data (Blue) and their WFOD transform (Red) on real ECG data from BIDMC Congestive Heart Failure database [8]

3. DATA AND EXPERIMENTAL SETUP

To test the classification performance of the proposed WFOD feature, a publicly available motion artifact anomaly contaminated ECG dataset with three conditions [9], which is obtained from Physionet website [10], is taken. In this dataset, the ECG data are collected from a total of 31 subjects while standing, walking, and jumping to observe the contamination of the data with motion artifacts. Here, the standing case is considered the control, the walking case refers to the low anomaly condition, and the jumping case reflects the high anomaly condition. The sampling rate of the data is 500Hz, and each case consists of 8 seconds of measurements. There are three offsets of electrode patches, namely 0, 45, and 90 degrees, and they are considered as the repetitive measures of the same subject in this study. Hereby, the experimental conditions for classification are generated in pairs, i.e., standing-walking, standing- jumping, and walking-jumping cases. Therefore, the classifications are performed independently on those pairs. Therefore, the data prepared for the classification part comprises 18 instances of measurement, consisting of 9 measurements per group in binary classification.

On the other hand, the features in this study are selected from the statistical moments of the data, which are mean, variance, skewness, and kurtosis. Here, the WFOD feature is applied as an additional feature to the pair of those features on half of the experimental conditions to observe the effects of the WFOD feature in terms of classification performances. The number of features is needed to be limited to a maximum of three due to the curse of dimensionality [11]. Therefore, the experiments are performed by using the pairs of such statistical moments.

RESULTS AND DISCUSSION

The classification accuracy results of four different classifiers for each individual pair of experiments are provided in the tables. Here, Table 1 refers to the experiment, standing and walking pair, i.e., control group and low anomaly group. Table 2 is for the standing and jumping pair, i.e., control group and high anomaly group. Finally, Table 3 represents the results for the walking and jumping pair, i.e., low anomaly group and high anomaly group.

In all analyses, we implement the well-known classification approaches, namely, tree algorithm, linear discriminant analysis (LDA), K-nearest neighbour approach (K-NN), and naive Bayes methods while comparing the performance of different feature pairs.

According to Table 1, WFOD feature has no significant change in the accuracies of the classifiers, where it either decreases or increases the accuracies by 1% to 6% with no consistency. Hence, it can be inferred that the WFOD feature has no significant effect on the discrimination of the control and low anomaly groups. Moreover, comparing the classifiers, the tree classifier has mostly better classification accuracies than other classifier algorithms.

TABLE I. Classification accuracy of statistical feature pairs with and without WFOD feature for standing - walking experiment under distinct classification methods.

Feature Pairs	Tree	LDA	K-NN	Naive Bayes
Mean + Variance	0.58	0.60	0.52	0.48
Mean + Variance + WFOD	0.57	0.61	0.51	0.49
Mean + Skewness	0.67	0.48	0.61	0.51
Mean + Skewness + WFOD	0.66	0.48	0.63	0.51
Mean + Kurtosis	0.73	0.53	0.56	0.49
Mean + Kurtosis + WFOD	0.67	0.52	0.62	0.50
Variance + Skewness	0.67	0.52	0.61	0.50
Variance + Skewness + WFOD	0.64	0.51	0.60	0.51
Variance + Kurtosis	0.73	0.58	0.56	0.53
Variance + Kurtosis + WFOD	0.72	0.56	0.63	0.47
Skewness + Kurtosis	0.70	0.52	0.67	0.60
Skewness + Kurtosis + WFOD	0.67	0.51	0.64	0.53
WFOD	0.66	0.51	0.46	0.50

TABLE II. Classification accuracy of statistical feature pairs with and without WFOD feature for standing - jumping experiment under distinct classification methods.

Feature Pairs	Tree	LDA	K-NN	Naive Bayes
Mean + Variance	0.75	0.57	0.73	0.56
Mean + Variance + WFOD	0.84	0.72	0.68	0.62
Mean + Skewness	0.87	0.73	0.71	0.68
Mean + Skewness + WFOD	0.8	0.80	0.76	0.66
Mean + Kurtosis	.87	0.66	0.61	0.58
Mean + Kurtosis + WFOD	0.8	0.80	0.70	0.62
Variance + Skewness	0.86	0.71	0.75	0.72
Variance + Skewness + WFOD	0.87	0.81	0.73	0.72
Variance + Kurtosis	0.87	0.8	0.68	0.62
Variance + Kurtosis + WFOD	0.83	0.78	0.73	0.67
Skewness + Kurtosis	0.86	0.7	0.76	0.76
Skewness + Kurtosis + WFOD	0.89	.80	0.75	0.76
WFOD	0.81	0.6	0.67	0.66

TABLE III. Classification accuracy of statistical feature pairs with and without WFOD feature for walking - jumping, experiment under distinct classification methods.

Feature Pairs	Tree	LDA	K-NN	Naive Bayes
Mean + Variance	0.78	0.66	0.65	0.53
Mean + Variance + WFOD	0.90	0.71	0.6	0.63
Mean + Skewness	0.77	0.69	0.69	0.66
Mean + Skewness + WFOD	0.90	0.65	0.73	0.71
Mean + Kurtosis	0.84	0.70	0.69	0.65
Mean + Kurtosis + WFOD	0.89	0.66	0.73	0.67
Variance + Skewness	0.77	0.67	0.67	0.66
Variance + Skewness + WFOD	0.92	0.67	0.69	0.66
Variance + Kurtosis	0.81	0.69	0.59	0.66
Variance + Kurtosis + WFOD	0.93	0.68	0.69	0.63
Skewness + Kurtosis	0.68	0.71	0.65	0.71
Skewness + Kurtosis + WFOD	0.93	0.69	0.70	0.70
WFOD	0.87	0.62	0.76	0.60

Table 2 refers to the comparison of the control and high anomaly groups, where the difference between the groups is expected to be highest among the pairs of groups. Here, it can be observed that the WFOD feature mostly provides an improvement in the classification accuracy from 0% up to 14%. However, there are a few cases in that WFOD

reduces the classification accuracy slightly up to 4%. For this pair of groups, again, the tree classifier provides better overall results than other classifiers.

Finally, Table 3 refers to the results of discriminating the low and high anomaly groups.

According to these results, the WFOD algorithm has a superior performance with the improvement of the accuracy by up to 25% in contribution to the discrimination of the groups. It can be observed that the WFOD feature performance is the best in this pair of groups. Moreover, similar to the previous pairs of groups, the tree classifier has the best assessment compared to the other classifiers.

CONCLUSION

In this study, a new hybrid feature, namely the WFOD feature, which is based on the data's time and frequency domain characteristics is proposed. The performance of the underlying feature is tested with an experimental setup, compared to the statistical moments used as features in a classification study. The results imply that the WFOD feature generally improves the classification accuracies of ECG data with different groups of anomalies. The effect of the WFOD feature is highest when the classification of low and high anomalies indicates an improvement up to 25% in classification accuracy. Moreover, the unimodal classification with the WFOD feature also provides a good classification rate up to 87% for low and high anomaly groups. Hereby, we consider that the current results of the study are promising, and some further modifications to the WFOD algorithm may even improve the performance of the WFOD feature in the classification of the time series data. Hence, as future work, we think to improve the WFOD algorithm and to test it on different data modalities. On the other hand, the comparison of classifiers yields that the tree classifier provides the best accuracy values among the list of classifiers. Accordingly, in the extension of this study, especially, with the applications on biomedical data, it is found that the tree classifier provides the highest accuracy values. Therefore, we consider to use this specific method as the main classifier in order to obtain the optimal pair of classifiers versus feature extraction approaches in biomedical data analyses.

REFERENCES

1. Cressie N 2015 Statistics for spatial data (John Wiley & Sons)
2. Akoglu L, Tong H and Koutra D 2015 Data mining and knowledge discovery 29 626–688
3. Hastie T, Tibshirani R and Friedman J 2009 The elements of statistical learning: data mining, inference, and prediction (Springer Science & Business Media)
4. Phinyomark A, Phukpattaranont P and Limsakul C 2012 Expert systems with applications 39 7420–7431
5. Kropf M, Hayn D and Schreier G 2017 2017 Computing in Cardiology (CinC) (IEEE) pp 1
6. Islam M K, Rastegarnia A and Yang Z 2016 Neurophysiologie Clinique/Clinical Neurophysiology 46 287–305
7. Erkuş E C and Purutçuoğlu V 2021 European Journal of Operational Research 291 560–57
8. Baim D S, Colucci W S, Monrad E S, Smith H S, Wright R F, Lanoue A, Gauthier D F, Ransil B J, Grossman W and Braunwald E 1986 Journal of the American College of Cardiology 7 661–670
9. Behravan V, Glover N E, Farry R, Chiang P Y and Shoaib M 2015 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN) (IEEE) pp 1–6
10. Goldberger A L, Amaral L A, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 circulation 101 e215–e220
11. Indyk P and Motwani R 1998 Proceedings of the thirtieth annual ACM symposium on Theory of computing pp 604–613