A Study on the Effects of Window Size on Electrocardiogram Signal Quality Classification

Tanatorn Tanantong
Rangsit University, Thailand, tanatorn.t@rsu.ac.th

ABSTRACT

The sliding window-based method is one of the most used method for automatic Electrocardiogram (ECG) signal quality classification. Based on this method, ECG signals are generally divided into small segments depending on a window size and these segments are then used in another classification process, e.g., feature extraction. The segmentation step is necessary and important for signal classification and signal segments with different window sizes can directly affect the performance of classification. However, in signal quality classification, the window size is often randomly selected and further analysis on the most appropriate window sizes is thus required. In this paper, an extensive investigation of the effects of window size on signal quality classification is presented. A set of statistical-amplitude-based features widely used in the literature was extracted based on 10 different window sizes, ranging from 1 to 10 seconds. To construct signal quality classification models, four well-known machine learning techniques, i.e., Decision Tree, Multilayer Perceptron, *k*-Nearest Neighbor, and Naïve Bayes, were employed. The performance of the quality classification models was validated on an ECG dataset collected using wireless sensors from 20 volunteers while performing routine activities, e.g., sitting, walking, and jogging. The evaluation results obtained from four machine-learning classifiers demonstrated that the performance of signal quality classification using window sizes of 5 and 7 seconds were good compared with other sizes.

Keywords: Electrocardiogram (ECG) signal, ECG signal quality classification, wireless monitoring systems, signal segmentation, window size.

I INTRODUCTION

Electrocardiogram (ECG) signals describe electrical activities of the human heart and they are primarily used in diagnosis and treatment of cardiovascular diseases (Catalano, 1993; Hampton, 2008; Norman, 1992). However, in data acquisition, especially in continuous monitoring, ECG signals are often corrupted by several types of noises, e.g., motion artifact, power line interference, and baseline drift noises (Clifford, Azuaje, & McSharry, 2006). Such noises considerably affect the quality of ECG signals and lead to high false cardiac alarm rates in intensive care units (Schmid, Goepfert, & Reuter, 2013). Therefore, the assessment of the signal quality is also an important process and required for continuous ECG monitoring.

Several researchers have addressed issues related to assessing quality levels of ECG signals and have proposed an automatic approach for signal quality classification. In 2011, the Computing in Cardiology Challenge (Silva, Moody, & Celi, 2011) was arranged by PhysioNet, aiming to find some effective and efficient methods for classifying quality levels of ECG signals captured using mobile phones. The PhysioNet has also publicly provided an ECG dataset for evaluating methods of signal quality classification. The dataset consists of 2000 ECG signal recordings, 10 seconds long each, collected using mobile phones. Each signal recording was manually annotated by experts with three quality levels, “Acceptable”, “Unacceptable”, and “Indeterminate”. Based on this ECG dataset, several studies proposed automatic approaches for ECG signal quality classification using different techniques, for example, combination of rule-based and machine-learning-based methods (Kuzilek, Huptych, Chudacek, Spilka, & Lhotska, 2011), Ensembles of Decision Trees (Zaunseder, Huhle, & Malberg, 2011), a threshold-based rule (Hayn, Jammerbund, & Schreier, 2012; Johannesen & Galeotti, 2012). However, further investigation on the effects of window size on signal quality classification and further experiments on ECG signals continuously captured while subjects are performing a routine daily activity were required.

Studies on the effects of window size for signal classification has been also addressed many research area, e.g. detecting embolic signals using the Fast Fourier Transform (FFT) (Aydin, 2000), analyzing electromyography signals (Thongpanja, 2013), and classifying acceleration signals for human activity recognition (Banos, Galvez, Damas, Pomares, & Rojas, 2014). These studies demonstrated the importance of analyzing the window size in the signal processing and motivated further investigation on the window size impacts on classifying the quality of ECG signals.

This paper presents an extensive study on the effects of window size on signal quality classification. ECG signals acquired using wireless devices from 20 volunteers while performing routine activities were used. Based on a sliding window technique, statistical-amplitude-based feature were extracted from ECG signals relying on a defined window size. Four machine learning algorithms, i.e., Decision Tree, Multilayer Perceptron, *k*-Nearest Neighbor, and Naïve Bayes, were employed to construct classification models. In order to investigate the effects of window size on signal quality classification, each classification model was evaluated using a different set of window sizes. The rest of this paper is organized as follows: Section 2 provides related works. Section 3 describes the materials and methods used for automatic classifying quality levels of ECG signals. Section 4 reports evaluation results and discusses the effects of window size on signal quality classification. Section 5 presents conclusions.

II RELATED WORKS

A. ECG Signal Quality Classificaiton

A combination of a rule-based method and a machine learning-based method (Kuzilek, et al., 2011) was proposed for ECG signal quality classification. A set of noise detection rules and a Support Vector Machine (SVM) classifier were employed for calculating a quality score of each signal recording based on statistical values of signal amplitudes and time-lagged covariance matrices. Using the signal quality scores determined from rules and SVM, an accuracy of 83.6% was achieved.

An automatic method based on Ensembles of Decision Trees (EDTs) for ECG signal quality classification was presented (Zaunseder, et al., 2011). In order to construct a EDTs classifier, frequency-domain features base on high frequency (45-250 Hz) and low frequency (0-0.5 Hz) noises in ECG signals, were used. The proposed method yielded an accuracy of 90.4%.

An algorithm for determining the quality of ECG signals (Johannesen & Galeotti, 2012) was developed, consisting of two steps: (1) exclusion of signal recordings with ECG-lead connection issues, using QRS complex information and (2) determination of signal quality levels of each recording, relying on noise type information. The two-step algorithm provided an accuracy of 90.0%.

Four quality measures based on empty lead, spike detection, lead-crossing point, and QRS-detection robustness criterion, were implemented in order to assess the quality of ECG signals (Hayn, et al., 2012). Using combination of these four measures, a good evaluation result was obtained for signal quality classification, with an accuracy of 91.6%.

In all above studies, the different methods for automatic signal quality classification were presented with high accuracy results on the PhysioNet ECG dataset. However, the effects of window size on signal quality classification and the experiments on ECG signals continuously captured while subjects are performing a routine daily activity were not yet investigated.

B. Window Size Effects in Signal Classifiation

An extensive study on the effects of window size in analysis and detection of embolic signals using the fast Fourier transform (FFT) was reported (Aydin, 2000). The embolic signals were acquired using a commercial Doppler ultrasonic system, EME Pioneer TC4040, with a frequency of 2 MHz. Based on the FFT technique, six dissimilar window sizes, 2.2, 4.4, 8.9, 17.9, 35.8, and 71.6 milliseconds, were employed for evaluating the effects of window size. The evaluation results showed that the FFT window sizes of 8.9 and 17.9 milliseconds were mostly suitable for detecting embolic signals.

For electromyography (EMG) signal processing, an investigation on impacts of window size on analyzing surface EMG signals (Thongpanja, 2013) was presented. In data collection, EMG signals were captured from 6 volunteers while they are performing 5 levels of lifting objects, weighed between 1 to 5 kilograms, and are performing a 5-second movement of elbow flexion and extension. Using six different window sizes, i.e., 125, 250, 375, 500, 750, and 1,000 milliseconds, EMG signals were segmented using a sliding window technique. For evaluation, the Modified Reverse Arrangement (MRA) test was adopted in order to assess the stationarity of EMG signal segments. The results demonstrated that 375 and 125 milliseconds were optimal values of window size for analyzing surface EMG signals for static and dynamic contractions, respectively.

In (Banos et al., 2014), a comprehensive study on the effects of window size on automatic human activity classification using acceleration signals was presented. In data acquisition, acceleration signals were captured from 17 subjects while they were performing 33 activities. Using different window sizes ranging from 0.25 to 7 seconds (in steps of 0.25 second) and a non-overlapping sliding window technique, three sets of statistical features including mean and standard deviation were extracted. Four machine learning algorithms, i.e., Decision Tree, *k*-Nearest Neighbors, Naïve Bayes, and Nearest Centroid Classifier, were employed to construct activity classification models. In order to evaluate the performance of each model, 10-fold cross validation and an F_1score measure were used. From the obtained results, the highest performance of activity recognition were achieved when using window sizes of 1 and 2 seconds.

All the studies mentioned above addressed the importance of selecting an optimal window size for analyzing and classifying different signals, i.e., embolic signals (Aydin, 2000), electromyography signals (Thongpanja, 2013), acceleration signals (Banos, et al., 2014). These studies motivated further investigation of the effects of window size on ECG signal quality classification.

III MATERIALS AND METHODS

A. Data Acquisition

In this study, ECG signals captured using wireless Body Sensor Networks (Yang, 2006) from 20 healthy volunteers, i.e., 10 young and 10 elderly volunteers, were used. The 10 young volunteers (7 males and 3 females, aged between 27-44 years) were asked to perform 16 daily activities, five times each, including standing, walking upstairs, up and down movement of both arms, and jogging. The 10 elderly volunteers (2 males and 8 females, aged between 57-71 years) were asked to perform 7 daily activities, five times each, e.g., sitting, lying, and walking. Lead-II configuration (Barill, 2005), which is usually applied for monitoring patients' ECG signals in Intensive Care Units, and a sampling rate of 100 Hz were employed for acquiring the signal recordings.

B. Signal Segmentaion and Annotation

For signal segmentation, a non-overlap sliding window method (Banos, et al., 2014) was adopted in this study. According to a report from the Association for Advancement of Medical Instrumentation (AAMI, 2002), an abnormal case in ECG signals should be reported within 10 seconds. This motivated time-period number was used as the maximum size value of ECG segments for classifying quality levels. In order to investigate the effects of the window size, 10 different window sizes, range of 1 to 10 second in steps of 1 second, were considered. ECG signals were divided into small portions based on the different window sizes as illustrated in Figure 1.

To annotate ECG signals with quality labels, the signal quality classification scheme (G.D. Clifford, Behar, Li, & Rezek, 2012) was applied. This scheme was used in several studies, focused mainly on classifying quality levels of ECG signals (Joachim, Julien, Qiao, & Gari, 2013; Li & Rajagopalan, 2014; Tanantong, Nantajeewarawat, & Thiemjarus, 2015). In this study, two suggested quality levels, "Low-quality" and "High-quality", were used for labelling entire ECG signals. Low-quality signals are the signals that are contaminated with high levels of noises and cannot be confidentially used for a physician's diagnosis. High-quality signals are the signals that are noiseless or contaminated with some little noises. In addition, for high-quality signals, the significant ECG signal components, e.g., P, Q, R, S, and T waves, must be completely identified. Table 1 shows the proportion of signal segments to quality levels in 10 different window sizes.

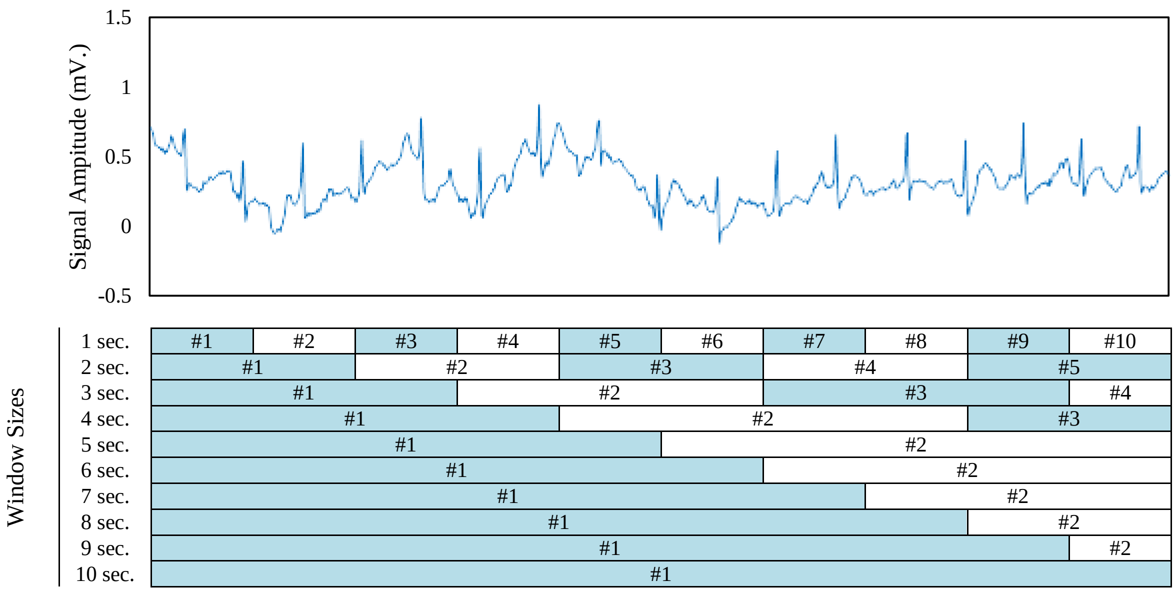

Figure 1. ECG Signals During Jogging (Top) and Examples of Signal Segments with Diverse Window Sizes (Bottom)

Table 1. No. of Segments in Each Window Size

Window Size (Second)	No. of Signal Segment		
	Low Quality	High Quality	Total
1	1850 (12.26%)	13237 (87.74%)	15087
2	944 (12.57%)	6567 (87.43%)	7511
3	598 (11.98%)	4395 (88.02%)	4993
4	460 (12.32%)	3273 (87.68%)	3733
5	374 (12.58%)	2600 (87.42%)	2974
6	301 (12.19%)	2168 (87.81%)	2469
7	266 (12.59%)	1847 (87.41%)	2113
8	221 (12.02%)	1618 (87.98%)	1839
9	200 (12.26%)	1431 (87.74%)	1631
10	185 (12.69%)	1273 (87.31%)	1458

C. Feature Extraction and Signal Quality Classificaiton

Based on 10 different window sizes, four statistical-amplitude-based features, i.e., mean, variance, slope, and difference between maximum and minimum values of ECG signal amplitudes in each segment, were extracted. The number of all obtained features can be determined by the multiplication between the number of features and the number of segments in the defined window sizes (Referring to Table 1). Such features were also widely employed in several studies on ECG signal quality classification (Chudacek, Zach, Kuzilek, Spilka, & Lhotska, 2011; Johannesen & Galeotti, 2012; Kuzilek et al., 2011; Schumm, Arnrich, & Troster, 2012). To construct the signal quality classification model, four widely known machine learning techniques (Witten, Frank, & Hall, 2005), i.e., Decision Tree, Multilayer Perceptron, *k*-Nearest Neighbor, and Naïve Bayes, were applied. In this study, the WEKA open-source data mining tool (Bouckaert et al., 2010) were utilized for the implementations of these machine learning techniques.

D. Performance Measures

The performance of the ECG signal quality classification is measured using four statistical measures, i.e., Sensitivity (*SEN*), Specificity (*SPE*), Selectivity (*SEL*), and Accuracy (*ACC*). These measures are given by:

$$SEN = \frac{TP}{TP+FN} \times 100\%$$

$$SPE = \frac{TN}{TN+FP} \times 100\%$$

$$SEL = \frac{TP}{TP+FP} \times 100\%$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%$$

where *TP* (True Positive) and *TN* (True Negative) are the number of signal segments accurately predicted as “Low Quality” and “High Quality”, respectively. *FP* (False Positive) and *FN* (False Negative) are the number of signal segments inaccurately predicted as “Low Quality” and “High Quality”, respectively.

For avoiding the effects of data unbalancing (Sokolova, 2009), the $F_1 score$ is additionally employed to evaluate the performance of the signal quality classification. This measure is a combination of Sensitivity and Selectivity measures, also known as Recall and Precision in text classification evaluation. It is defined as follows:

$$F_1 score = 2 \times \frac{SEN \times SEL}{SEN + SEL} \times 100\%$$

IV RESULTS AND DISCUSSION

For evaluating the signal quality classification, a 10-fold cross validation technique (Witten, et al., 2005) was employed. Table 2 illustrates the performance results of signal quality classification using dissimilar window sizes (1 to 10 seconds) and four classification algorithms, i.e., Decision Tree (DT), Multilayer Perceptron (MLP), *k*-Nearest Neighbor (*k*-NN), and Naïve Bayes (NB). The overall performance results were between 68.78% and 88.65% for sensitivity, between 96.02% and 98.89% for specificity, between 72.14% and 90.06% for selectivity, and between 93.01% and 96.37% for accuracy.

The DT classifier yielded the highest accuracy of 96.21% using a 7-second window size and the highest sensitivity of 98.89% for 10 seconds. Utilizing an 8-second window size, the top specificity and selectivity of 98.89% and 89.41 were obtained, respectively. For MLP with a 5-second window size, the best accuracy, specificity, and selectivity of 96.37%, 98.73%, and 90.06% were obtained, respectively. The top sensitivity value was 81.58% for a window size of 7 seconds. Using *k*-NN ($k = 3$) and a 5-second window size, the classifier gained the highest accuracy and sensitivity of 79.68% and 96%, respectively. The top specificity and selectivity of 98.43% and 87.82% were obtained when using a window size of 7 seconds. The NB classifier with 7-second window size provided the highest accuracy, specificity, selectivity of 95.08%, 97.51%, and 81.89%, respectively. For a 5-second window size, NB achieved the maximum sensitivity of 79.41%. The performance results obtained from each algorithm demonstrate that the window size has different effects on the signal quality classification.

Table 2. Performance Comparision of Classification Algorithms Using 10 different Window Sizes

Algorithm	Measure	Window Sizes (Second)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Decision Tree (DT)	**SEN**	74.97%	81.36%	78.76%	77.17%	83.96%	79.73%	81.95%	68.78%	86.00%	**88.65%**
	SPE	98.45%	97.69%	98.57%	97.98%	97.54%	98.15%	98.27%	**98.89%**	97.00%	97.01%
	SEL	87.12%	83.48%	88.20%	84.32%	83.07%	85.71%	87.20%	**89.41%**	80.00%	81.19%
	ACC	95.57%	95.63%	96.19%	95.42%	95.83%	95.91%	**96.21%**	95.27%	95.65%	95.95%
Multilayer Perceptron (MLP)	**SEN**	74.54%	77.12%	79.77%	77.83%	79.95%	79.07%	**81.58%**	73.30%	82.00%	79.46%
	SPE	98.69%	98.52%	98.54%	97.86%	**98.73%**	98.57%	98.16%	98.58%	98.25%	98.11%
	SEL	88.80%	88.24%	88.17%	83.64%	**90.06%**	88.48%	86.45%	87.57%	86.77%	85.96%
	ACC	95.72%	95.83%	96.29%	95.39%	**96.37%**	96.19%	96.07%	95.54%	96.26%	95.75%
***k*-Nearest Neighbor (*k*-NN)**	**SEN**	72.11%	73.31%	77.26%	75.87%	**79.68%**	78.07%	78.57%	78.73%	75.00%	77.84%
	SPE	98.35%	98.20%	98.50%	97.74%	98.35%	98.02%	**98.43%**	97.84%	97.90%	98.04%
	SEL	85.90%	85.43%	87.50%	82.51%	87.39%	84.53%	**87.82%**	83.25%	83.33%	85.21%
	ACC	95.13%	95.07%	95.95%	95.04%	**96.00%**	95.59%	95.93%	95.54%	95.10%	95.47%
Naïve Bayes (NB)	**SEN**	70.11%	72.03%	75.75%	75.65%	**79.41%**	75.75%	78.20%	76.02%	75.50%	77.84%
	SPE	96.22%	96.03%	96.52%	96.36%	97.00%	96.54%	**97.51%**	96.97%	96.02%	96.94%
	SEL	72.14%	72.26%	74.75%	74.52%	79.20%	75.25%	**81.89%**	77.42%	72.60%	78.69%
	ACC	93.01%	93.01%	94.03%	93.81%	94.79%	94.01%	**95.08%**	94.45%	93.50%	94.51%

Note: SEN = Sensitivity, SPE = Specificity, SEL = Selectivity, ACC = Accuracy

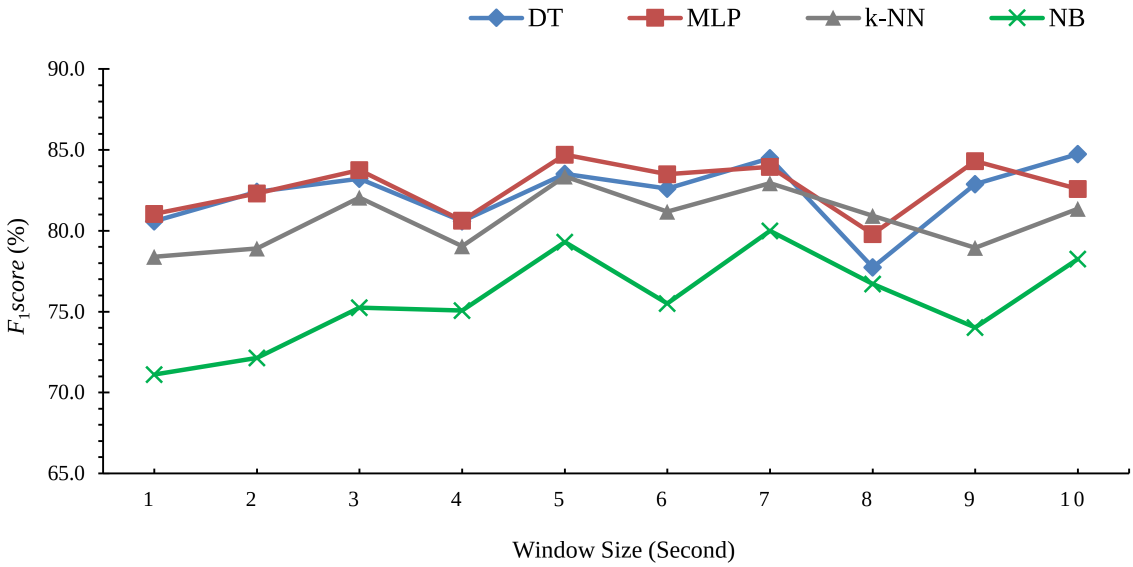

Figure 2. Effects of the Window Size on Signal Quality Classification Performance (*F_1Score*)

performance. Using window sizes of 5 and 7 seconds, the good classification results are mostly achieved (e.g., MLP, *k*-NN, and NB). Conversely, for all algorithms, the low performance results are always obtained for window sizes lower than 5 seconds.

However, in order to avoid data unbalancing impacts, the F_1*score* measure was additionally used for evaluating the performance of each classification model. Figure 2 illustrates the F_1*score* results of signal quality classification using four machine learning algorithms and different window sizes. The overall results were between 71.11% and 84.75%. For MLP and *k*-NN (k = 3) using a 5-second window size, the maximum performance results were obtained, which were 84.70% and 83.36, respectively. Using DT, the best and second-best results, 84.75% and 84.50%, were achieved using window sizes of 10 and 7 seconds, respectively. NB showed the top performance result with an F_1*score* equal to 80% for 7 seconds.

These obtained results show that the most performance of classifying ECG signal quality levels was increased when the window size was increased until 7 seconds. For example, the NB provided the minimum results (71.11%) for 1 second. It yielded the second-best and best results (79.31% and 80%) using window sizes of 5 and 7 seconds, respectively. Although some algorithms, DT and MLP, achieved the good performance results for a couple of higher window sizes (e.g., 9 and 10 seconds), such results were not much different from the experimental results using the 5-second and 7-second window sizes. Moreover, utilizing the window sizes between 1 to 4 seconds and an 8-second window size should not be suggested for ECG signal quality classification.

V CONCLUSION

A comprehensive study on the effects of window size on ECG signal quality classification has been proposed. In this study, several machine-learning-based methods techniques and signal-amplitude-based features, which were widely employed in previous works, were used for constructing signal quality classification models. For investigating the window size effects, 10 different window sizes, ranging from 1 to 10 seconds, were considered in the experiments. As demonstrated by the evaluation results, the suitable window sizes were 5 and 7 seconds and the use of sizes between 1 to 4 seconds were not suggested for classifying the quality of ECG signals.

ACKNOWLEDGMENT

This work was supported by the Research Institute of Rangsit University and the Thailand Research Fund (TRF), under Royal Golden Jubilee Ph.D. Program.

REFERENCES

AAMI. (2002). Cardiac monitors, heart rate meters, and alarms, american national standard (Revision of ANSI/AAMI EC13:2002). Association for the Advancement of Medical Instrumentation, Arlington, VA, Tehnical Report.

Aydin, N., & Markus, H.S. (2000). Optimization of processing parameters for the analysis and detection of embolic signals. European Journal of Ultrasound, 12(1), 69-79.

Banos, O., Galvez, J.-M., Damas, M., Pomares, H., & Rojas, I. (2014). Window size impact in human activity recognition. Sensors, 14(4), 6474-6499.

Barill, T. P. (2005). The six second ECG: A practical guidebook to basic ECG interpretation. Palm Springs, California, USA: SkillStat Learning Inc.

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKA-Experiences with a Java Open-Source Project. J. Mach. Learn. Res., 11, 2533-2541.

Catalano, J. T. (1993). Guide to ECG analysis: J. B. Lippincott.

Chudacek, V., Zach, L., Kuzilek, J., Spilka, J., & Lhotska, L. (2011). Simple scoring system for ECG quality assessment on android platform. Proceedings of the Computing in Cardiology Conference, Hangzhou, China.

Clifford, G. D., Azuaje, F., & McSharry, P. (2006). ECG statistics, noise, artifacts, and missing data Advanced methods and tools for ECG data analysis (pp. 55-99). London: Artech House.

Clifford, G. D., Behar, J., Li, Q., & Rezek, I. (2012). Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. Physiol. Meas., 33(9), 1419-1433.

Hampton, J. R. (2008). 150 ECG problems: Elsevier.

Hayn, D., Jammerbund, B., & Schreier, G. (2012). QRS detection based ECG quality assessment. Physiol. Meas., 33(9), 1449-1461.

Joachim, B., Julien, Q., Qiao, L., & Gari, D. C. (2013). ECG Signal Quality During Arrhythmia and Its Application to False Alarm Reduction. IEEE Trans. Biomed. Eng., 60(6), 1660-1666.

Johannesen, L., & Galeotti, L. (2012). Automatic ECG quality scoring methodology: mimicking human annotators. Physiol. Meas., 33(9), 1479-1489.

Kuzilek, J., Huptych, M., Chudacek, V., Spilka, J., & Lhotska, L. (2011). Data driven approach to ECG signal quality assessment using multistep SVM classification. Proceedings of the Computing in Cardiology Conference, Hangzhou, China.

Li, Q., & Rajagopalan, C. C., G. (2014). A machine learning approach to multi-level ECG signal quality classification. Computer Methods and Programs in Biomedicine, 117(3), 435-447.

Norman, A. E. (1992). 12 lead ECG interpretation a self-teaching manual: McGRAW-HILL.

Schmid, F., Goepfert, M. S., & Reuter, D. A. (2013). Patient monitoring alarms in the ICU and in the operating room. Crit. Care., 17(2), 216-222.

Schumm, J., Arnrich, B., & Troster, G. (2012). ECG monitoring in an airplane seat: appraising the signal quality. IEEE Pervasive Comput., 11(4), 28-34.

Silva, I., Moody, G. B., & Celi, L. (2011). Improving the quality of ECGs collected using mobile phones : the PhysioNet/Computing in cardiology challenge 2011. Proceedings of the Computing in Cardiology Conference, Hangzhou, China.

Sokolova, M., & Lapalma, G. (2009). A systematic analysis of performance measures for classification tasks. Inf. Process Manag., 45(4), 427-437.

Tanantong, T., Nantajeewarawat, E., & Thiemjarus, S. (2015). False alarm reduction in BSN-Based cardiac monitoring using signal quality and activity type information. Sensors, 15(2), 3952-3974.

Thongpanja, S., Phinyomark, A., Quaine, F., Laurillau, Y., Wongkittisuksa, B., Limsakul, C., & Phukpattaranont, P. (2013). Effects of window size and contraction types on the stationarity of biceps brachii muscle EMG signals. Proceedings of the 7th International Convention on Rehabilitation Engineering and Assistive Technology, Gyeonggi, South Korea.

Witten, I. H., Frank, E., & Hall, M. A. (2005). Data Mining: Practical machine learning tools and techniques, 3rd Edition. Burlington: Morgan Kaufmann/Elsevier.

Yang, G. Z. (2006). Body sensor networks. London: Springer-Verlag.

Zaunseder, S., Huhle, R., & Malberg, H. (2011). CinC challenge-assessing the usability of ECG by ensemble decision trees. Proceedings of the Computing in Cardiology Conference, Hangzhou, China.