Advances in Social Science, Education and Humanities Research, volume 483
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

On the Problem of Class Imbalance in the Recognition of Electrocardiograms

Marat Bogdanov*
Department of Computational Mathematics and Cybernetics
Ufa State Aviation Technical University
Department of Applied Informatics
M. Akmullah named after Bashkir State Pedagogical University
Ufa City, Russia
bogdanov_marat@mail.ru

Nikolai Oskin
Siberian Telemetric Company
Moscow City, Russia
nonik2@mail.ru

Irina Dumchikova
Department of Applied Informatics
M. Akmullah named after Bashkir State Pedagogical University
Ufa City, Russia
redfoxufa@gmail.com

***Abstract*—The paper is about the problem of class imbalance in the diagnosis of diseases of the cardiovascular system using recognition of electrocardiograms. Under researching two oversampling approaches were compared. Complete cardiocycles (600 points) were used as features. In the first case, the bootstrap method was used. For recognition, a Multylayer Perceptron neural network was used. To solve the problem, significant computational resources and high costs of computer time were required. In the second case, cardiocycles were converted into images oversampled by augmentation. The calculation time was reduced from two and a half hours to 15 minutes. In the third case, both approaches were combined, which reduced the computation time to three minutes. In all three cases, recognition accuracy exceeded 97%.**

Keywords—Machine Learning, Electrocardiogram, Deep Learning, Diagnoses of heart diseases

I. INTRODUCTION

Currently, one of the booming areas of computer science is data analysis. Big data has its own specifics. First of all, a huge amount of processed and stored information is very impressive. An example is the Large Hadron Collider. So, its detectors generate for about 1 petabyte of raw data per second. This data is processed by 500,000 processor cores. The processed data is stored on various warehouses. Each year, the volume of such information increases by 100 petabytes.

CMS (Compact Muon Solenoid) (Fig. 1) is a particle detector that is designed to see a wide range of particles and phenomena produced in high-energy collisions in the LHC. Like a cylindrical onion, different layers of detectors measure the different particles, and use this key data to build up a picture of events at the heart of the collision. It produce ~ 10 Petabytes/sec of information, 50 Petabytes of data per year.

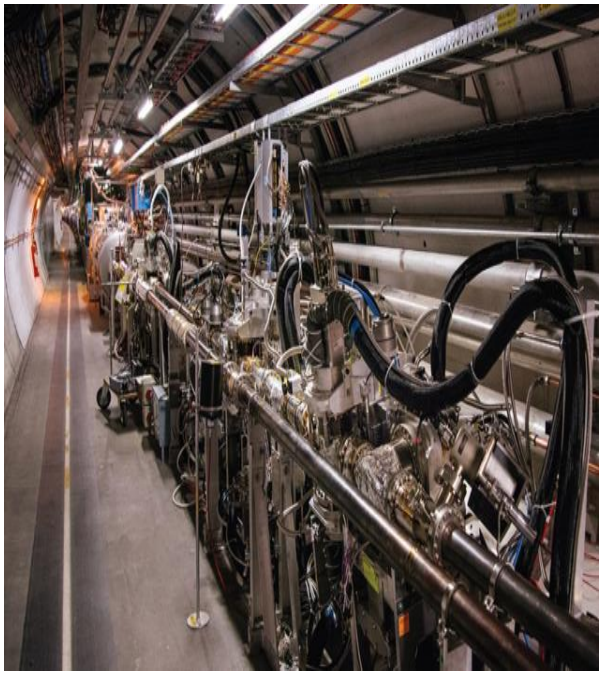

Fig. 1. CMS detector.

ATLAS (A Toroidal LHC ApparatuS) (Fig. 2) [1] is the largest, general-purpose particle detector experiment at the Large Hadron Collider (LHC), a particle accelerator at CERN (the European Organization for Nuclear Research) in Switzerland. The experiment is designed to take advantage of the unprecedented energy available at the LHC and observe phenomena that involve highly massive particles which were not observable using earlier lower-energy accelerators. It's generated 170 PB of data.

Copyright © 2020 The Authors. Published by Atlantis Press SARL.
This is an open access article distributed under the CC BY-NC 4.0 license -http://creativecommons.org/licenses/by-nc/4.0/.

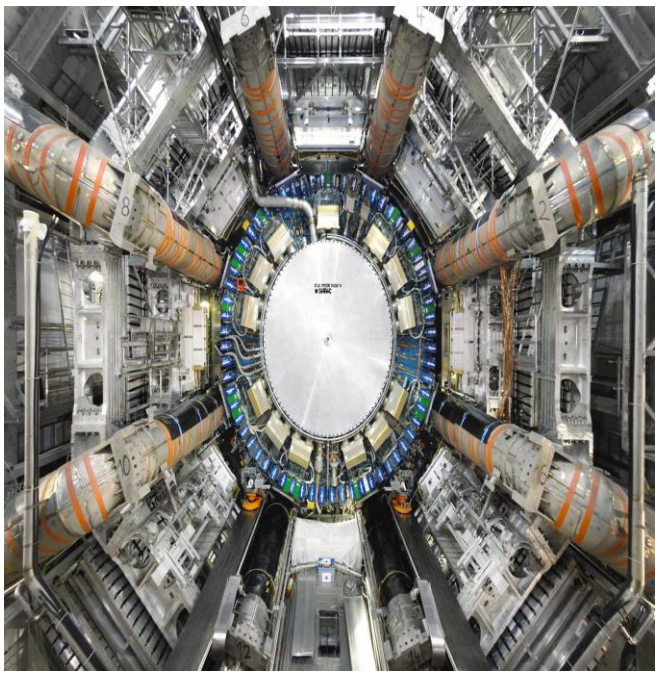

Fig. 2 Atlas detector.

The second feature of big data is the need for fast data transfer. In the case of the Large Hadron Collider, the data transfer rate from CERN to scientific centers located around the World is 6 gigabytes per second (600 terabytes per day). The third feature of big data is that events of interest to scientists usually occur very rarely. Scientists at CERN working with the Large Hadron Collider research the products of numerous proton collisions with each other. One interesting event takes place on 10^{13} routine events. In Russia, people talk about a needle in a haystack, in America people talking about finding a grain of sand on a beach with an area of 20 volleyball fields.

A similar situation has also takes place with the American project Pan-STARRS (Panoramic Survey Telescope and Rapid Response System). This system includes several observatories equipped with telescopes with 1.4 gigapixel sensors. Each observatory scans of ¾ of the sky 3 times a month, forming a five-color image of 30 terabytes. Several petabytes of data are accumulated over the year. The mission's goal is to search and track celestial objects larger than 300 meters that can cross the Earth's orbit. It is necessary to simultaneously track about 80 million celestial objects, of which 2-3 asteroids can pose a real threat to the Earth [1].

The problem of finding a needle in a haystack is also often encountered in the analysis of various medical images, for example, in the diagnosis of oncology. Here we are faced with the fourth feature of big data - rare events can be very important.

Currently the methods of machine and deep learning are widespread in the analysis of data. The first one includes the support vector machines or the random forest classifier, and the second one includes various neural networks. Using these methods, classification problems can be solved when the trained model can predict the class label, that is, determine to which category the studied object belongs.

When teaching machine or deep learning models, it is desirable that the number of objects in each class will be approximately the same. Otherwise, an overfitting situation may occure when the model recognizes major classes well and ignores minor classes. To overcome the problem of class imbalance several approaches are proposed. One of them is based on cost sensitive learning: assigning a high cost to misclassification of the minority class, and trying to minimize the overall cost. [2].

The second approach is to multiply data from minor classes (oversampling) or reduce the amount of data in the major classes (undersampling). Such a strategy has several implementations. The SMOTE approach (Synthetic Minority Over-sampling Technique) is best known [3].

II. Motivation and aim

Cardiovascular disease is the reason of a large number of deaths and cases of disability worldwide. An important tool for diagnosing cardiovascular diseases is electrocardiography. Currently, telemedicine services are actively developing, using machine learning methods to recognize various pathological conditions using electrocardiogram analysis. Usually Machine Learning methods works good if classes are balanced, that is, the sample size in each class should not vary greatly, otherwise there may be a problem of overfitting when minor classes are ignored due to the model's training in the major class. Minor classes are often very important, as they are often associated with severe cardiological dysfunctions that can lead to the death of the patient. This may explain the complexity of collecting relevant data. To train methods for classifying electrocardiograms, publicly available databases of biomedical signals are often used. An example is the Physikalisch-Technische Bundesanstalt (PTB) database [4], available at www.physionet.org [5]. This database contains the digitized electrocardiograms of 290 subjects belonging to 15 diagnostic classes. The sample is extremely unbalanced. A similar situation is quite common. For reliable results, class balancing is required. The proposed work will describe two approaches aimed at solving the problem of class imbalance.

III. Previous works

Various rhythm disturbances plays important role Among diseases of the cardiovascular system. Cardiologists have long been collecting of annotated electrocardiograms associated with arrhythmias. An example is the MIT-BIH database, a joint project of Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) and Massachussets Institute of Technology. The project has been developing since 1975. This database contains cardiocycles assigned to 16 classes, which AAMI (AAMI EC57: 1998) recommended combining into five groups depending on their physiological origin [2] (Table 1).

Table 1

Relationship betwen MIT-BIH cardiocycles and AAMI standards

AAMI classes	***MIT-BIH cardiocycles***
Non-ectopic beats (N)	Normal beats Left bundle branch block beats Right bundle branch block beats Nodal (junctional) escape beats Atrial escape beats
Supra ventricular ectopic beats (S)	Aberrated atrial premature beats Supraventricular premature beats Atrial premature Contraction Nodal (junctional) premature beats
Ventricular ectopic beats (V)	Ventricular flutter wave Ventricular escape beats Premature ventricular contraction
Fusion beats (F)	Fusion of ventricular and normal beat
Unknown beats (Q)	Paced beats Unclassifiable beats Fusion of paced and normal beats

The distribution of the number of cardiocycles by classes was as follows (Table 2)

Table 2

Total number of cardiocycles and types of cardiocycles

AAMI classes	***Number of cardiocycles***
N	3240
S	2506
V	2298
F	467
Q	1351

Kandala et al. performed researching in classification of various arrhythmias. The authors worked with the MIT-BIH database and used three methods for balancing classes in their work [6, 7].

1. Resampling

Matthew F. Dixon et al. developed the Over Sampling for Time Series Classification (OSTSC) tool for balancing time series using the Long Short-Term Memory (LSTM) neural network and Tensorflow technology. The authors obtained good results in terms of increasing recognition accuracy on a wide range of tasks from high-frequency trading to the classification of arrhythmias.

2. Oversampling technology for synthetic minor classes (SMOTE).

This method was proposed by Chawla. The method is intended to generate the "synthetic" minority samples instead of duplicating the samples. SMOTE works is as following manner:

• Selecting the k-nearest neighbors for a chosen feature vector based on the oversampling.

• Taking a difference between selected sample and its nearest neighbor of each.

• Multiply the difference vectors by a random number between 0 and 1.

• A new "synthetic" sample will be generated by adding this vector to the selected minority sample.

This approach generalizes the decision region of the minority class.

Sampling techniques used to solve the problems with the distribution of a dataset, sampling techniques involve artificially re-sampling the data set, it also known as data preprocessing method. Sampling can be achieved by two ways, Under-sampling (Fig. 3) the majority class, oversampling the minority class, or by combining over and undersampling techniques.

Fig. 3 explains under sampling procedure.

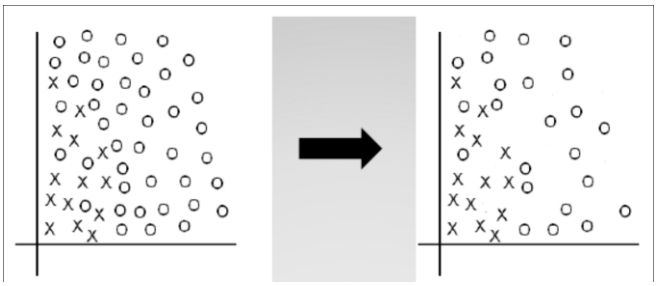

Fig.3. Randomly removes the majority sample. [8]

Fig. 4. explains oversampling procedure.

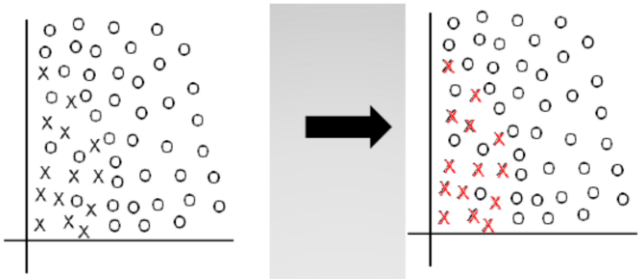

Fig.4. Replicate the minority class samples [8]

3. Data sampling based on distribution of data.

Imbalanced data sets severely suffer from class oversampling and disjuncts. This issue is addressed by preprocessing through distribution based balancing. We should do following operations:

• Find the prior probability distribution of each feature f i , i = 1, 2, . . ., r given class label c j as p(f i /c =c j)

• Sample new b instances of each class with the prior probability distribution knowledge. In this way, artificially generated balancing set is formed by sampling equal number of instances in each class.

The authors obtained good results in increasing recognition accuracy on a wide range of tasks from high-frequency trading to the classification of arrhythmia.

IV. MATERIALS AND METHODS

We used the PTB database containing 549 samples of digitized electrocardiograms obtained from 290 subjects [4].(aged 17 to 87, mean 57.2; 209 men, mean age 55.5, and 81 women, mean age 61.6; ages were not recorded for 1 female and 14 male subjects). Each subject is represented by one to five records. There are no subjects numbered 124, 132, 134, or 161. Each record includes 15 simultaneously measured signals: the conventional 12 leads (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6) together with the 3 Frank lead ECGs (vx, vy, vz). Each signal is digitized at 1000 samples per second, with 16 bit resolution over a range of ± 16.384 mV. On special request to the contributors of the database, recordings may be available at sampling rates up to 10 KHz.

After signal preprocessing, features were classified. We used a complete cardiocycle (600 points) as features. The distribution of cardiocycles on classes is given in table 3.

Table 3.

Sample sizes related to different diagnostic classes in the PTB database

Diagnostic class	***Number of cardiocycles***
Bundle branch block	2356
Cardiomyopathy	2248
Dysrhythmia	1303
Healthy control	10591
Heart failure (NYHA 2)	37
Heart failure (NYHA 3)	55
Heart failure (NYHA 4)	38
Hypertrophy	992
Myocardial infarction	52124
Myocarditis	477
Palpitation	46
Stable angina	161
Unstable angina	38
Valvular heart disease	500
n/a	2313

Table 3 shows that the Myocardial infarction is a major class. If you do not oversample the remaining classes, then they will be very poorly recognized. To balance the classes, three approaches were used.

In addition, we used the European ST-T Database [5]. The European ST-T Database is intended to be used for evaluation of algorithms for analysis of ST and T-wave changes. This database consists of 90 annotated excerpts of ambulatory ECG recordings from 79 subjects. The subjects were 70 men aged 30 to 84, and 8 women aged 55 to 71. The database includes 367 episodes of ST segment change, and 401 episodes of T-wave change, with durations ranging from 30 seconds to several minutes, and peak displacements ranging from 100 microvolts to more than one millivolt. Each record is two hours in duration and contains two signals, each sampled at 250 samples per second with 12-bit resolution over a nominal 20 millivolt input range. The distribution of cardiocycles on classes is given in table 4.

Table 4.

Sample sizes related to different diagnostic classes in the European ST-T Database

Diagnostic class	**Number of cardiocycles**
1-vessel disease (LAD)	220
1-vessel disease (LCX)	20
1-vessel disease (RCA)	20
2-vessel disease	80
2-vessel disease (LAD LCX)	20
2-vessel disease (LCX RCA)	20
2-vessel disease (RCA LAD)	60
3-vessel disease	80
3-vessel disease and LCA main stem	20
Angina pectoris	40
Anterior myocardial infarction	20
Arterial hypertension	20
Chest pain	100
Coronary artery by-pass graft	140
Coronary artery disease	20
Hyperkalemia	20
Inferior myocardial infarction	20
Infero-lateral myocardial infarction	20
Medications: diltiazem molsidomine	160
Medications: nifedipine beta-blockers	20
Medications: nitrates	20
Myocardial infarction	80
No coronary angiography	20
Non-Q myocardial infarction	160
Normal coronary arteries	100
Normal coronary artery vessels	20
Normal coronary vessels	180
Recorder type: Reynolds Tracker	20
Resting angina	20
Unknown coronary artery disease	20

As we can see the major classes are: 1-vessel disease (LAD), Coronary artery by-pass graft, Medications: diltiazem

molsidomine, Non-Q myocardial infarction and Normal coronary vessels.

In addition to this, we used a St Petersburg INCART 12-lead Arrhythmia Database [6].

This database consists of 75 annotated recordings extracted from 32 Holter records. Each record is 30 minutes long and contains 12 standard leads, each sampled at 257 Hz, with gains varying from 250 to 1100 analog-to-digital converter units per millivolt. Gains for each record are specified in its .hea file. The reference annotation files contain over 175,000 beat annotations in all. The original records were collected from patients undergoing tests for coronary artery disease (17 men and 15 women, aged 18-80; mean age: 58). None of the patients had pacemakers; most had ventricular ectopic beats. In selecting records to be included in the database, preference was given to subjects with ECGs consistent with ischemia, coronary artery disease, conduction abnormalities, and arrhythmias. The distribution of cardiocycles on classes is given in table 5.

Table 5.

Sample sizes related to different diagnostic classes in the St Petersburg INCART 12-lead Arrhythmia Database

Diagnostic class	Number of cardiocycles
Acute MI	100
Coronary artery disease	60
Earlier MI	120
PVC	460
Sinus node dysfunction	180
Transient ischemic attack	340
VT	40
tachycardia	20
ventricular bigeminy	60

As we can see the major classes are: PVC, Transient ischemic attack.

a. Oversampling by bootstrap method

Oversamling was performed according to Odile Pons [9]. We have written a script in Python. The volume of each class was brought to the size of the major class (52124 samples). As a result, the feature matrix began to contain 781860 rows. To classify electrocardiograms, a Multiplayer Perceptron neural network was used. Due to the large amount of computation, we had to use the google colab service using a tensor accelerator. The calculation time was two and a half hours. Recognition accuracy exceeded 97 percent.

b. Oversampling by image augmentation method

Image augmentation artificially creates training images through different ways of processing or combination of multiple processing, such as random rotation, shifts, shear and flips, etc. For example we can get the initial image and perform series of transformations by rotations and horizontal shif. On the figure 1 we can see the initial image (Fig.5).

Fig. 5. The initial image. [11]

On the Fig. 6. we can see augmented images.

Fig. 6. series of transformations of initial image [11].

In the second case, we converted 600-point cardiocycles into 24 * 25 images (Fig. 7).

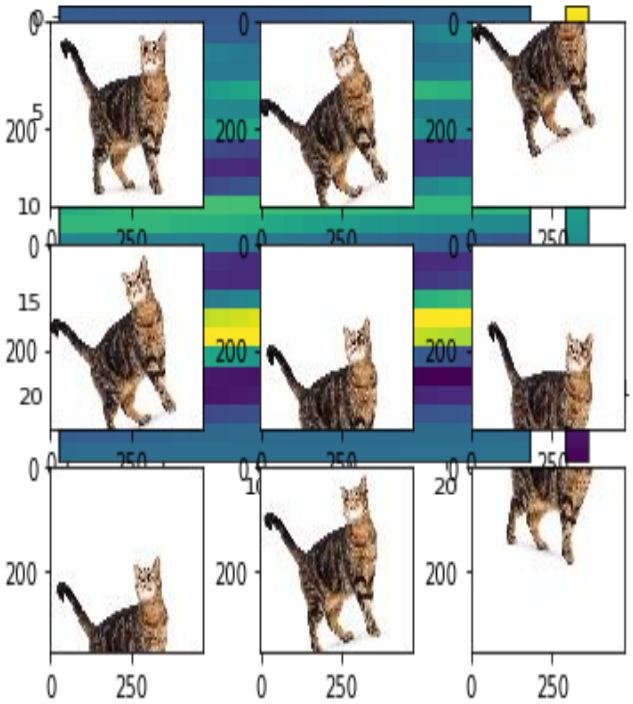

Fig. 7. Cardiocycle converted to 24 * 25 image.

Every diagnosis of heart diseases has own specific image pattern (Fig. 8).

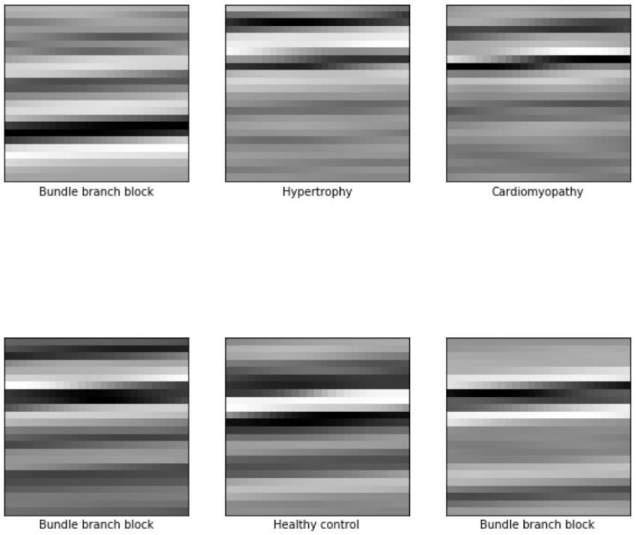

Fig. 8. Image patterns of heart diseases.

Further, for oversampling the samples, we used the image augmentation method (Suki Lau. Image Augmentation for Deep Learning [10]), fitting the number of images in each class to the size of the major class. We used image warping, rotation, changing the illumination, sprinkling with salt and pepper. Image recognition was carried out using convolution neural network also on google colab service. Recognition time decreased to 15 minutes. Recognition accuracy also exceeded 97%.

c. Hybrid approach

In the third case, the oversampling of cardiocycles by the bootstrap method was first performed, then all 781860 rows of the feature matrix were converted into images and recognized ng a convolution neural network. Recognition time was reduced to three minutes.

V. Conclusion

Under researching, two oversampling approaches were compared. Complete cardiocycles (600 points) were used as features. In the first case, the bootstrap method was used. For recognition, a Multylayer Perceptron neural network was used. To solve the problem, significant computational resources and high costs of computer time were required. In the second case, cardiocycles were converted into images oversampled by augmentation. The calculation time was reduced from two and a half hours to 15 minutes. In the third case, both approaches were combined, which reduced the computation time to three minutes. In all three cases, recognition accuracy exceeded 97%.

Acknowledgment

The reported study was funded by RFBR according to the research project No 19-07-00780

References

[1] *The Pan-STARRS1 data archive home page. https://panstarrs.stsci.edu/*

[2] R.C. Prati, G.E. Batista, M.C. Monard, A study with class imbalance and random sampling for a decision tree learning system, in: IFIP International Conference on Artificial Intelligence in Theory and Practice, Springer, 2008, pp. 131–140.

[3] IN.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[4] Bousseljot, R.; Kreiseler, D.; Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik, Band 40, Ergänzungsband 1 (1995) S 317.

[5] Web site of National Institute of Health. https://www.physionet.org/content/ptbdb/1.0.0

[6] Taddei A Distante G Emdin M Pisani P Moody GB Zeelenberg C Marchesi C. The European ST-T Database: standard for evaluating systems for the analysis of ST -T changes in ambulatory electrocardiography. European Heart Journal 13: 1164-1172 (1992).

[7] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.

[8] Kandala N.V.P.S. Rajesh, Ravindra Dhuli. Classification of imbalanced ECG beats using re-sampling techniques and AdaBoost ensemble classifier. Biomedical Signal Processing and Control 41 (2018) 242–254

[9] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50.

[10] T. Mar, S. Zaunseder, J.P. Martínez, M. Llamedo, R. Poll, Optimization of ECG classification by means of feature selection, IEEE Trans. Biomed. Eng. 58 (8) (2011) 2168–2177.