COMP6208: Advanced Machine Learning

Brent De Hauwere[1], Matthew De Vries[2], Ilias Kazantzidis[3], Dimitris Mallios[4]

Team name: Dr. Christiaan Barnards Assistants
[1]bdh1g19, [2]mrdv1n19, [3]ik3n19, [4]dm1n19

1 Introduction

Electrocardiogram (ECG) analysis has been established at the core of cardiovascular pathology diagnosis since its development in the twentieth century [14]. The main problem with manual analysis of ECG signals, similar to many other time-series data, lies in difficulty of detecting and categorising different waveforms and morphology in the signal. For a human, this task is both extensively time-consuming and prone to errors [10].

One of the most common heart diseases which, according to the NHS, affects more than two million people every year in UK, is arrhythmia. The rhythm of the heart is controlled by electrical impulse signals. ECG monitoring can capture the cardiac signals in order to analyse abnormalities. However, the monitoring of the cardiac signals using ECG is not flawless. The signals can be interfered with additional noise (e.g muscle noise), which complicates the evaluation process and can sometimes mislead doctors into false diagnosis affecting the patient's longevity. The open access to ECG databases has led to the development of many methods and approaches for machine learning (ML) ECG arrhythmia classification over the last decade. Figure 1 illustrates a normal single ECG signal wave which provides useful information about the functioning of the heart.

In this report we will analyse the data as well as discuss possible pre-processing and feature extraction techniques which may be used before training a model.

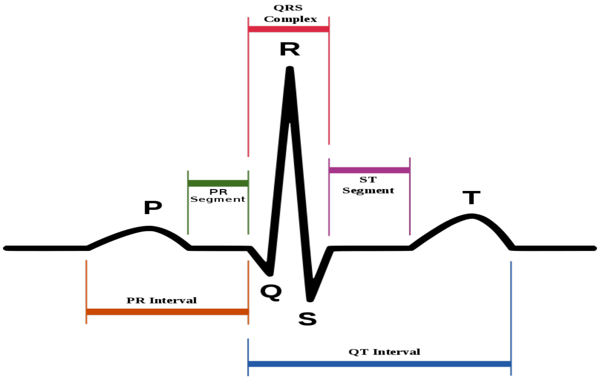


Figure 1: Schematic representation of normal sinus rhythm showing normal wave, segments, and intervals [Source: Wikipedia]

2 The Dataset

2.1 Unprocessed Dataset

The MIT-BIH dataset consists of 48 half-hour excerpts of ECG recordings from 47 different subjects recorded at the sampling rate of 360 Hz. Each beat is annotated by at least two cardiologists. This consists of 109,446 annotations of 5 categories [15]. A subsection of the labels are [4]:

- V — Ventricular Ectopic Beat (VEB): a ventricular premature beat, or a ventricular escape beat.
- F – Fusion Beat: a fusion of a ventricular and normal beat.
- Q — Paced Beat: a fusion of a paced and normal beat, or a beat that cannot be classified.
- S — Supraventricular Ectopic Beat (SVEB): an atrial or nodal (junctional) pre-mature or escape beat, or an aberrant atrial premature beat.
- N — Normal: any beat that does not fall into the V, F, Q, or S categories.

Figure 2 shows a heat map for each category in order to look for intra-category correlations. In this dataset, all beats were padded to have a total of 187 samples. Therefore, we only show the first 70 samples because the colour spectrum of the graphs would be too compressed to be able to infer information otherwise. The colour of a pixel in the figure indicates how many of the beats in that respective category had a certain amplitude at a given sample number (read: point in time). In ML algorithms, these strong intra-category correlations are what we aim to exploit.

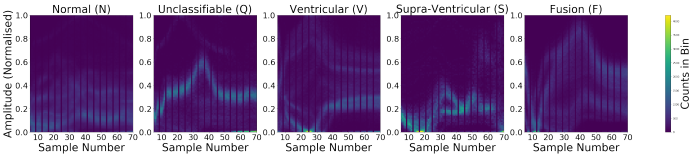


Figure 2: Heatmap for Every Category

2.2 Statistical Properties

Several statistics can be extracted from ECG signals. We will investigate the mean and variance of the number of peaks in a signal, as well as the peak prominence. The prominence of a peak is a measure of how much a peak stands out from the surrounding baseline of the signal. It is defined as the vertical distance between the peak and its lowest contour line. Table 1 gives an overview of the statistics for each category. We additionally calculated the average highest and top-five highest prominence value. Note that we did not take into account all prominence values because it was not informative as the majority was insignificant - diminishing the average. From this table, we can conclude that, except for category S, the mean number of peaks is not very distinguishable (which may be due to the noise). The same conclusion can be made for the average highest prominence. On the other hand, the variance in the number of peaks and the average top-five prominence's vary strongly depending on the category.

	N	Q	V	S	F
No. Peaks Mean	24.45	24.83	25.02	13.08	22.43
No. Peaks Var	110.6	145.2	176.6	44	77.94
Top-1 Prom.	0.64	0.64	0.64	0.56	0.70
Top-5 Prom. Mean	0.65	0.24	0.48	0.52	0.60

Table 1: Intra-Category Features: Peak Mean, Peak Variance, Top-1 Prominence, and Top-5 Prominence Mean

3 Related Work

Figure 3 illustrates the most famous techniques used for the preparation of data in signal processing (specifically the ECG) ML problems. The techniques used in this research will be described in more detail in §4. Now we will give an overview of the whole process, excluding classification which we will face in the next handin.

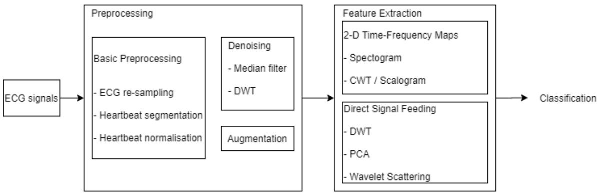


Figure 3: Common pre-processing and feature extraction methods for ECG signals

Initially, some basic pre-processing has to be applied to each of the ECG signals. ECG re-sampling (specifically down-sampling) is mostly used for compression purposes and it is equivalent to the familiar process in images of resizing them to a smaller resolution. Heartbeat segmentation (and detection) is necessary for obtaining the samples. The annotations of doctors from the dataset occur at each R peak (see Figure 1). Heartbeat normalisation of voltage (y axis) is also useful to keep the signal independent of the original amplitude.

Heartbeat signals contain noise from different resources such as movement of the patients, electrodes and instrument uncertainty. Median filtering and more effectively Discrete Wavelet Transform (DWT) [6, 8] can reduce this noise and at the same time retain the useful information of the signal (P, Q, R, S, T areas). MIT-BIH arrhythmia dataset is highly imbalanced. Consequently, in order to avoid bias towards that class, techniques such as data augmentation (§4.2) can be used [11, 5].

Regarding feature extraction, two approaches are usually followed. The first is to create 2-D time-frequency representations and tackle the signal as an image in the form of a Spectrograms (discussed in §4.5.2) which is then fed to the classifier. Scalograms are the absolute values of the Continuous Wavelet Transform (CWT) which can be additionally combined with transfer learning[1].

The second approach is to feed signals directly to the classifier, usually after some dimensionality reduction method. In literature PCA and DWT [6, 8, 2] are used extensively in pre-processing; the details are discussed in sections 4.3 and 4.4. Another powerful method is the Wavelet Scattering (Invariant Scattering Convolutional Network), which is a framework of three steps in correspondence with CNNs - Convolution (Wavelets), Nonlinearity (Modulus), Averaging (Scaling Function) - which provides low-variance features without losing critical information [1].

4 Data Exploration

Having seen the available tools from previous section, Figure 4 shows the techniques used and implemented in this research. We start with the pre-processing block.

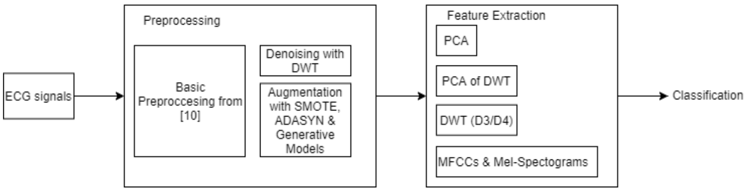


Figure 4: Our approach to pre-processing and feature extraction

4.1 Basic Pre-processing: Adapted Dataset

The dataset we adapted and built on[2], comes from [10] and the basic pre-processing applied can be easily understood from Figure 5. The details of this approach are discussed thoroughly in [10]. The main reasons why we consider it appropriate and adapt it are the following:

1. The total number of samples is the same as number of annotations by doctors given in the original dataset (i.e. exactly 109,446).

2. By downsampling to 125 Hz (from 360 Hz) and zero padding, all the extracted heartbeats (observations) consist of 187 samples [3] which is manageable for experiments with minimally affect on classification results.

3. This basic pre-processing step does not really affect the final results. As long as all the P, Q, R, S, T parts are present in one heartbeat, and the same parts are not present in the previous or in the next sample, the information will be the same whether we segment the recording by setting the R peaks to the centre or to the edges.

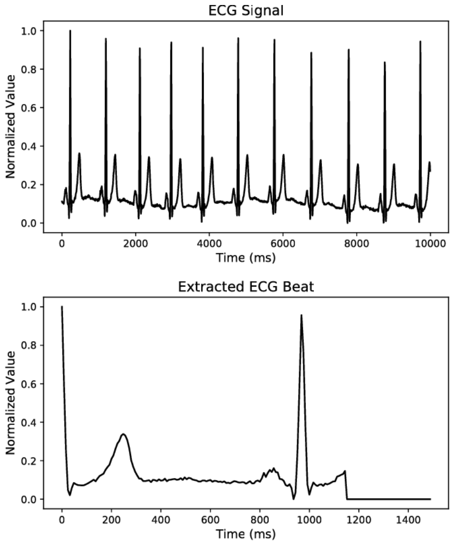


Figure 5: Segmentation of heartbeat (reproduced from [10])

4.2 Data Imbalance

ML algorithms are built to minimise errors. Since the probability of instances belonging to the majority class is significantly high

[1] A highly instructive example is presented from mathworks: `https://uk.mathworks.com/help/wavelet/examples/classify-time-series-using-wavelet-analysis-and-deep-learning.html`

[2] obtained on kaggle: ECG Heartbeat Categorization Dataset, `https://www.kaggle.com/shayanfazeli/heartbeat`

[3] samples in terms of signal processing - not ML

in an imbalanced data set, the algorithms are much more likely to classify new observations to the majority class. With respect to classification of myocardial arrhythmia, the cost of a false negative is much larger than a false positive. There are several articles addressing the issue with imbalanced data [11, 12]. ML algorithms penalise false positives and false negatives equally. A way to counter this is to modify the algorithm itself to boost predictive performance on minority class. This can be executed through either recognition-based learning [5] or cost-sensitive learning [7]. Another approach consists of re-sampling the data in order to mitigate the effect caused by class imbalance. The re-sampling approach has gained popular acceptance among practitioners as it is more flexible and allows for the use of latest algorithms. The two most common techniques are over-sampling and under-sampling. Over-sampling increases the number of minority class members in the training set. The advantage of over-sampling is that no information from the original training set is lost, as all observations from the minority and majority classes are kept. On the other hand, it is prone to overfitting. Under-sampling, on contrary to over-sampling, aims to reduce the number of majority samples to balance the class distribution. Since it is removing observations from the original data set, it might discard useful information. Finally, one can use generative techniques in order to generate a larger data set such a General Adversarial Networks or WaveNet. Our dataset is highly imbalanced with a distribution [0.066, 0.007, 0.025, 0.073, 0.828] for [V, F, S, Q, N] respectively. We have used two over-sampling methods in this work.

4.2.1 Synthetic Minority Over Sampling Technique (SMOTE)

Chawla et al. [3] developed an algorithm which applies a K nearest neighbours (KNN) approach where it selects KNN, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbours and computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from it.

4.2.2 Adaptive Synthetic (ADASYN)

Haibo et al. [9] presented a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples.

4.3 Denoising with DWT

Wavelet Transform is a powerful tool used both in signals and images for denoising and feature extraction. ECG signals are non-stationary, that is, the signal changes form in time, thus the frequency spectrum changes with respect to time as well. Consequently a transform like Fourier can not capture the time changes efficiently. Short Time Fourier Transform (STFT) trades off some frequency information for time information. However, Wavelet Transform can capture both. The reason, conceptually, is that a wavelet (a zero-mean finite oscillation) is scaled and shifted through a signal and compared with the latter, in order to obtain the correlation coefficients. High scale (extension of the wavelet) captures the low frequencies of the signal and low scale (contraction of the wavelet) conversely. Consequently, the full information from both frequency and time is encoded in these correlation coefficients. The preceding procedure deflects CWT.

In DWT, the scaling and shifting is discrete (scaling uses the base of 2) and the length of coefficients is approximately the same with the input signal (less memory). The exact process is illustrated in Figure 6 where high and low pass filtering is applied sequentially, and downsampling is acceptable due to Nyquist criterion. Di (detail coefficients) and Ai (approximation coefficients) represent the coefficients that encode the high and low frequency information at each level of decomposition. [4]

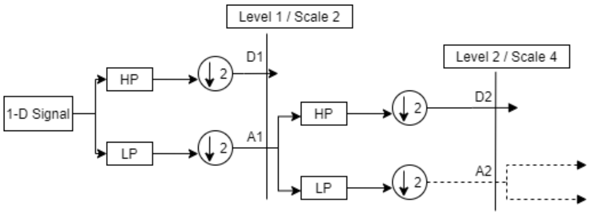


Figure 6: DWT decomposition

Specifically, denoising can be achieved by suitably thresholding (Universal and Soft thresholding was used) and shrinking/removing detail coefficients which include the noise towards 0. Then by reconstructing the signal from the remaining coefficients the noise will be reduced. We followed literature [6] which proposes 9 levels of decomposition and the Daubechies D6 wavelet. An example of the original and denoised signal is illustrated in Figure 7.

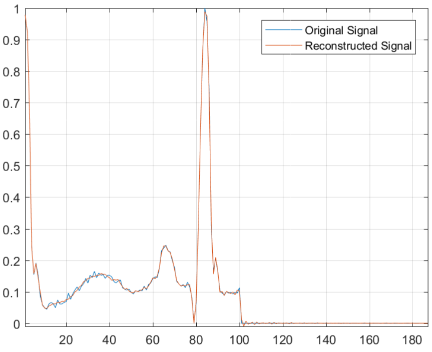


Figure 7: Denoising with DWT

Finally, it is worth mentioning that denoising takes time and is not a compulsory step if we aim for fast automatic classifiers.

4.4 Features from PCA and DWT

Principal Component Analysis (PCA) and DWT are predominantly used for feature extraction in ECG signals. They can be used independently or in conjunction (PCA of DWT).

Regarding only PCA, Figure 8 shows how for a specific number of eigenvalues (i) used, the area under the curve and to the right of i (directions ignored) is equal to the loss of energy. We realise that we do not need all 187 features. Hence, we decided to keep 100 eigenvalues (reducing dimensions almost to the half) and

[4] description and figure about dwt were adapted by `https://uk.mathworks.com/help/wavelet/ref/wavedec.html` , which toolbox was used for the decomposition

observed that the reconstructed signal is almost identical to the original one.

Concerning DWT, medical research on ECG has proposed different candidate parameters and features (type of wavelet, levels of decomposition, coefficients used as features). We choose the Meyer wavelet [13], whose shape is similar to the QRS complex (most representative part of each heartbeat[5]) so it can capture a lot of useful information. We also choose 4 levels of decomposition as proposed in most relevant researches. In total from DWT 189 coefficients are extracted: 12 A4 + 12 D4 + 24 D3 + 47 D2 + 94 D1 (Figure 6). After that, we follow two different directions:

1. Extract again 100 features with the PCA transform from all 189 DWT coefficients.

2. Keep the detail coefficients of the third (D3) and fourth (D4) layer (i.e. 12+24=36 coefficients in total).

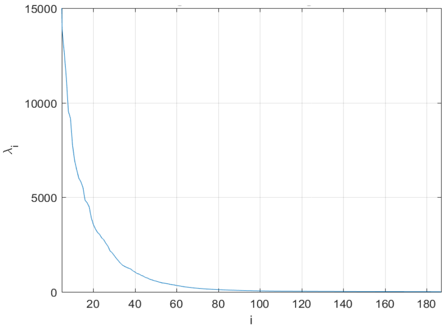


Figure 8: Eigenvalues of a random ECG Signal

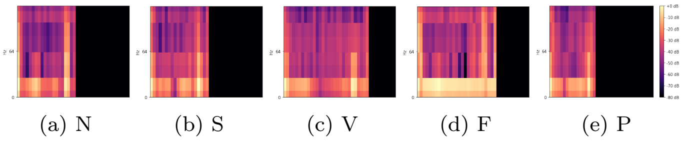
4.5 Frequency Based Features

4.5.1 Mel Frequency Cepstral Coefficients

One of the most famous feature representation methods in the field of sound and speech processing are the Mel-Frequency cepstral coefficients (MFCCs) [16]. The advantage of MFCCs relies on the fact that the frequency bands of the signal are distributed according to the Mel-scale. The Mel-scale is a non linear transformation method which is a good approximation of the human's auditory system instead of utilising a simple linear model. The previous statement holds from the fact that the bands in the Mel-scale are equally spaced and as a result provide a better representation of the sound signals. ECG is basically a 1D signal similar to a sound signal but captured at a much lower frequency. The dimensionality of the feature vector is 12 Mel-features per frame plus one energy feature from each frame resulting to 13 features. In this work we produce the first and the second derivatives from the MFCCs resulting to 39 features.

4.5.2 Mel-spectrograms

A different feature representation of the signal which is based on Mel-scale is the Mel-spectrograms. Spectrograms present the amplitude of signals at each frequency at a sequence of timesteps. Mel-Spectrograms are derived from the same algorithm as MFCC except for extracting the features after applying the Mel filter banks. The representation is a 2D binary image where the x-axis is the time and the y-axis is the frequency and each pixel has an intensity value representing the energy of the signal at each point in time at the given frequency. Audio signals are sampled in high frequency, from 16 kHz to 44 kHz using a 512-2048 size FFT window with 50% overlap. In our problem using the rule of three and from experiment we use a FFT window size of 10 with 50% overlap providing a 25Hz resolution regarding that our signals have low frequency. Figure 9 presents the Mel-spectrograms from each individual heartbeat category signals. The vivid yellow colour presents the high intensity of the signal in the corresponding frequencies while the darker purple colour presents a lower intensity respectively. The dark colours on the right of each sample occurred from the zero padding.



(a) N (b) S (c) V (d) F (e) P

Figure 9: Mel spectrogram examples from each heartbeat category.

[5] previously we had chosen db6 because the QRS complex was not a special landmark of the heartbeat in terms of denoising

References

[1] Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Transactions on Signal Processing*, 62(16):4114–4128, 2014.

[2] Angelo Catalani, Aris Anagnostopoulos, and Ioannis Chatzigiannakis. Arrhythmia classification from ecg signals. 2013.

[3] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.

[4] Gari D. Clifford and M. Oefinger. Ecg acquisition , storage , transmission , and representation. 2007.

[5] Chris Drummond and Robert C. Holte. Severe class imbalance: Why better algorithms aren't the answer. In João Gama, Rui Camacho, Pavel B. Brazdil, Alípio Mário Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005*, pages 539–546, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[6] Fatin A Elhaj, Naomie Salim, Arief R Harris, Tan Tian Swee, and Taqwa Ahmed. Arrhythmia recognition and classification using combined linear and nonlinear features of ecg signals. *Computer methods and programs in biomedicine*, 127:52–63, 2016.

[7] Charles Elkan. The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle*, 1, 05 2001.

[8] Nahit Emanet. Ecg beat classification by using discrete wavelet transform and random forest algorithm. In *2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, pages 1–4. IEEE, 2009.

[9] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, June 2008.

[10] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 443–444. IEEE, 2018.

[11] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[12] Joffrey L. Leevy, Taghi M. Khoshgoftaar, Richard A. Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):42, 2018.

[13] Y. Meyer. *Ondelettes et opérateurs: Ondelettes.* Actualités mathématiques. Hermann, 1990.

[14] Alfaras Miquel, Soriano Miguel C., and Ortín Silvia. A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection. *Frontiers in Physics*, 7:103, 2019.

[15] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[16] Douglas O'Shaughnessy. *Speech communication: human and machine.* Addison-Wesley, 1987.