MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190182

Electrocardiogram Beat-Classification Based on a ResNet Network

Cláudia Brito, Ana Machado, António Sousa

HASLab, INESC TEC & University of Minho, Braga, Portugal

Abstract

When dealing with electrocardiography (ECG) the main focus relies on the classification of the heart's electric activity and deep learning has been proving its value over the years classifying the heartbeats, exhibiting great performance when doing so. Following these assumptions, we propose a deep learning model based on a ResNet architecture with convolutional 1D layers to classify the beats into one of the 4 classes: normal, atrial premature contraction, premature ventricular contraction and others. Experimental results with MIT-BIH Arrhythmia Database confirmed that the model is able to perform well, obtaining an accuracy of 96% when using stochastic gradient descent (SGD) and 83% when using adaptive moment estimation (Adam), SGD also obtained F1-scores over 90% for the four classes proposed. A larger dataset was created and tested as unforeseen data for the trained model, proving that new tests should be done to improve the accuracy of it.

Keywords:
Electrocardiogram, Deep Learning, Arrhythmia

Introduction

Real-time monitoring has become one of the most important and clinically relevant tasks in medical settings, yet one of the most repetitive and tiresome tasks is the analysis of 24-hour ECG records. One of the ways to automate this long task is to convert this process into a real-time process with the automatic classification of the heart rate and with this, the classification of arrhythmias.

Arrhythmias are the most common diagnoses in this medical area and are composed of electrical changes that cause the normal heart rhythm to change. These changes can cause the heart to beat faster (tachycardia), slower (bradycardia), or at an irregular beat. Arrhythmias are widely classified, and their classification may depend on the factors described above and, where they occur, ventricles or atria. These changes can even lead to sudden death from stroke or cause other types of damage because of the inability of the heart to pump enough blood into the body and consequently cause damage to the brain, heart, or other organs [1,2].

To this extent, the importance of monitoring systems increases with the extra goal of improving patient care as well as the speed with which such care is provided.

Over the years, several approaches to this topic have been built. Going from the detection of the R-peak with high precision [3] to the creation of frameworks for this subject [4,5], the heart rhythm has been deeply studied. More recently, researchers have jumped from the classification of the heart beats with traditional methods to machine learning methods and even further, deep learning methods [6,7].

It is palpable the need to not only disclose the heart rhythm as tachycardia, bradycardia or irregular rhythms but to classify each of the beats into a defined category. Although being a deafening task, several works have presented great results when performing this task. Many researchers have spent their time around this subject using several methodologies as Roopa, C. and Harish B. [7] and Salem, A. et al [8] made notice in their survey. Support vector machines [9], genetic algorithms [10], rough set theory, and hidden markov models [11] and more lately neural networks [12–19], several other works mixing different methodologies have also been proposed [20].

In the second semester of 2017, Rajpurkar et al. [21] proposed a ResNet architecture of 34 layers to classify ECG batches of 30 seconds. This work exceeded the performance of high qualified cardiologists in a dataset 500 times larger than the overall datasets and set the classification task a step further to the automated analysis.

On the other side, researchers have also been focusing their efforts in patient-specific methodologies. The neural networks are trained individually for each patient allowing to classify future holters from the same patient [22,23]. Both works show promising outcomes and present a basis for future studies.

We intend to propose a new deep learning model to classify three distinct types of heart beats (four different classes) while analysing differents perspectives from the related works hereby addressed. This paper presents an overview of the dataset used as well as the architecture of the deep learning model built, providing insights on how the model was trained, and the results obtained with prospects of future work to be done. The dataset created, from records obtained from a local hospital, provided new insights about the model. The results obtained for the MIT-BIH Arrhythmia Database were promising with the arrhythmia classification yielding 96% accuracy in the classification of each beat in each one of the four classes used. The same network was used in the larger dataset, being able to classify the arrhythmias with an accuracy of 81%.

Materials and Methods

Data Selection

The MIT-BIH Arrhythmia Database presents 48 records, where the last 23 records became online only in 2005. The 48 records were chosen from a set of over 4000 long-term Holters recorded in the laboratories of the late Boston's Beth Israel Hospital, now known as Beth Israel Medical Center, between 1975 and 1979. The first 23 records were chosen randomly from inpatients and the 25 other records were chosen from the same set yet to contain a diversity of uncommon but clinically important conditions. The two groups have different purposes since the first one is to serve as a representative sample of waveforms and artifacts which most of the arrhythmia detectors might encounter normally. While the second set of records presents more complex arrhythmias and other conduction abnormalities.

The signals were sampled at 360Hz and recorded with a two-channel recorder. The annotations were made based on a simple slope-sensitive QRS detector and by two cardiologists, who added additional beat labels missed by the QRS detector and changed all the labels of abnormal beats. Nevertheless, during the following years many records had their beats relabeled by users who reported errors in the annotations [24,25]. These records are composed by three files, an .hea format file, .dat format file and a .atr format file.

The new dataset proposed comprises 113 records from 24-hour holters from a local hospital. These records were sampled at 125Hz. This dataset presents 2172 hours of data, while the first dataset, comprises only 48 records with 30 minutes each making a total of 24 hours of data. While the first dataset, after some preprocessing and discarded beats ended up with 97737 beats on all 4 classes, each record from the local hospital presented 100000±20000 beats (Figure 1). The created dataset was first analysed by the system's software and corrected by a technician. Then, each signal's classification was validated by a cardiologist.

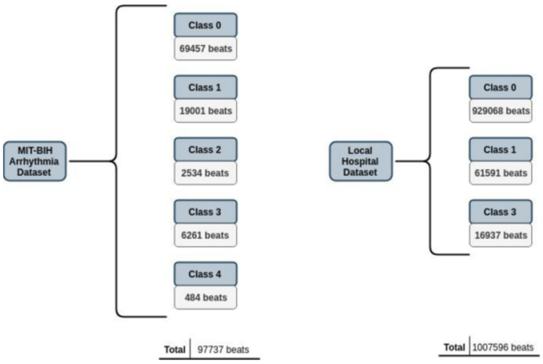

Figure 1 – Summary of Number of Beats of Each Class for Both Datasets

Later, it was disclosed the need to perform data augmentation in the classes proposed. The second and third classes were augmented since the number of beats belonging to these classes were low.

Data Treatment

The first step to analyze an ECG record begins with the load of the record and posterior filtering. These two processes were performed with the help of the CardIO library, an open-source Python library which was built to create "end-to-end machine learning models for deep research of electrocardiograms" [4].

The CardIO library relies in the WFDB package [26] to read and load files in the MIT-BIH format, this library can be executed via command line and also as a python library (the code is publicly available on Github). Using the capabilities of CardIO library, the signals were resampled to a frequency, of 125Hz. Then the signal was filtered, the filtering used in this preprocessing was based on several works that proved to be efficient [27-30]. Thus, it uses a band-pass filter, a finite impulse response filter (FIR filter [31]) which uses a frequency of 0.5Hz and 60Hz, in accordance with the theoretical foundations, yeat the new dataset was filtered with a frequency of 50 Hz and not 60Hz due to the specifications of the country's baseline wandering.

Nevertheless, in order to build the final datasets for training and testing, after these transformations, the ECG is sliced by beats and labeled with the annotations available creating the datasets for training and testing. The annotations were carefully reviewed and converted accordingly to the types of arrhythmias that were intended to classify. Therefore, the annotations were converted into four classes: normal beats as 0, other rhythms as 1, atrial premature contractions as 2 and premature ventricular contractions as 3, as seen in Table 1.

Table 1 – Representation of Each Type of Beat According to its Class

Type of Beat	Class
Normal beats	0
Other Rhythms	1
Atrial Premature Contractions	2
Premature Ventricular Contractions	3

The records were then split, based on the correct detection of the R-Peak and since it is sampled at 125Hz, it was decided to use a window with a size of 120 data points.

Model

We built a ResNet architecture (Figure 2 presents the high-level architecture of the network) based on Convolutional 1D layers for the classification task since this was the model that

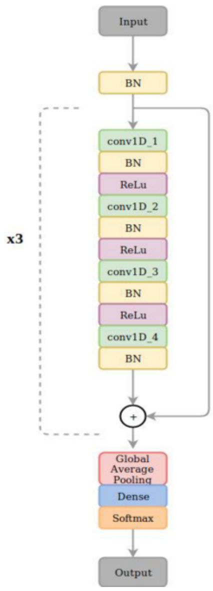

Figure 2 – Network Structure of the Model for the Classification Task. Overall the network contains 12 layers of convolution followed by a fully-connected layer and a Softmax.

surpassed the first tests performed. Initially, it was created a multilayer perceptron model and a version with 2D convolutional layers of the current model. It takes as input a time-series of 120 data points that represents the beat and outputs its label prediction. The model is composed by an initial input in a BatchNormalization layer, followed by four blocks of

Convolution1D layers (conv1D_1, conv1D_2, conv1D_3 and conv1D_4) with BatchNormalization, ReLU as activation function and a stride of 1. The conv1D_1, conv1D_2, conv1D_3 and conv1D_4 layers have a filter size of 8, 5, 3 and 1, respectively. The model could also use a Dropout schema, but considering the model complexity, we believe that we could obtain good results with BatchNormalization.

Evaluation Metrics

The effectiveness of a deep learning model for classification is usually measured by several parameters. Four of the most common are i) accuracy, ii) precision, and iii) recall, as well as the iv) F1-score [32–34]. These parameters are based on the results of true positives, true negatives, false positives and false negatives.

- True Positive (TP): correctly classified positive instances
- True Negative (TN): correctly classified negative instances
- False Positive (FP): incorrectly classified as positive
- False Negative (FN): correctly classified as negative

Accuracy

The Accuracy measures the rate of True Negatives and True Positives on all classified instances, being generically represented for binary classification by Equation 1. This result may induce in error or can hide important details since it does not distinguish between the number of correct labels of different classes. It can be understood as the number of correct predictions on top of the total number of prediction. The problems of misclassification may arise even though the accuracy results are high. In the medical field, the misdiagnose of an unhealthy subject may cost its life. When dealing with multi-classification, this formula reverts to the number of well predicted samples in relation to the total number of samples.

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP} \tag{1}$$

Precision

Precision (Equation 2) equalizes to the rate of true positive instances in the correctly classified instances.

$$\boldsymbol{Precision = \frac{TP}{TP + FP}} \tag{2}$$

Recall

Recall or sensitivity (Equation 3) gives higher scores when a high number of true positives is achieved while avoiding false negatives, this defines the true positive rate.

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \tag{3}$$

F1-Score

The F1-score is another general metric that is broadly used for evaluating these systems, the Equation 4 combines the precision and recall into a single number and is seen as one of the most reliable metrics for evaluating machine learning results.

$$F1_{score} = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4}$$

Experimental Results

The tests performed for training the model were optimized with SGD and Adam, both with a learning rate of 0.1. Other optimisers could have been used, however, we relied on SGD and Adam based on the related work and their broad use[35]. To understand the evolution of the models with the number of epochs, both systems were tested with 10, 25, 50, 100 and 150 epochs, the maximum number of epochs was defined after the first tests, where any test over 150 epochs demonstrated that the model stopped learning. As seen in Figure 3, the first dataset was randomly sliced in 80% for training and 20% for testing, with the training set being further separated into train and validation (65/35%).

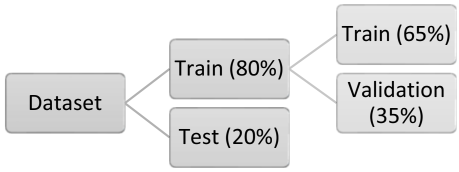

Figure 3 – Modeling the First Dataset for Training and Testing the Model

Several tests were performed for tuning the hyperparameters and adapting the learning rates to the behavior the model was exhibiting. Since this is a multi-classification task, the metrics for each class are calculated.

Figure 4 presents the overall accuracy of the models with a different number of epochs, a batch size of 2000 beats and as we are dealing with multi-classification, it used the categorical cross-entropy as loss function. The accuracy of the model reached 96% for the SGD optimizer with 150 epochs

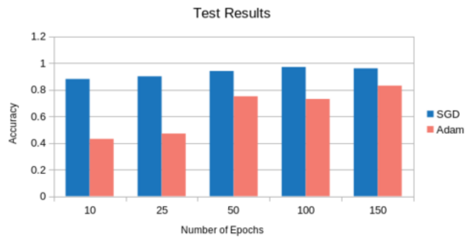

Figure 4 – Testing Results Regarding the Accuracy of the Model using Stochastic Gradient Descent and Adaptive Moment Estimation as Optimizers.

On the other hand, Figures 5 and 6 rely on the F1-score regarding the test results with the 20% of the dataset created.

The model behaves accordingly to what was expected, presenting good results for classifying all the instances initially defined. However, it can be noticed the higher results of the model optimized with stochastic gradient descent. These results are an effect of how both these optimizers work.

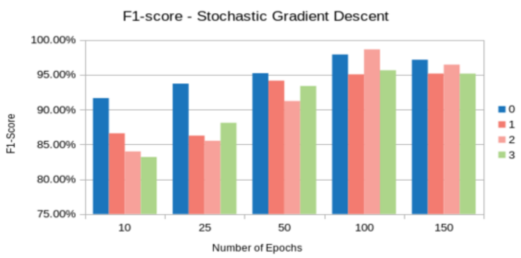

Figure 5 – Testing Results Regarding the F1-Score of the Model using Stochastic Gradient Descent as Optimizer

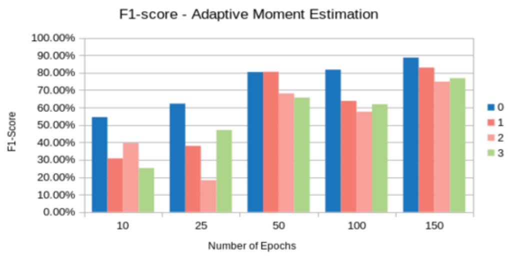

Figure 6 – Testing Results Regarding the F1-Score of the Model Using Adaptive Moment Estimation as Optimizer

The results obtained did not present overfitting, the error was minimal, and the test set obtained good results both in accuracy and F1-score. The model was able to classify all four classes and these results increased with the number of epochs.

When scrutinizing the two types of arrhythmia classified it is possible to disclose the reasons why atrial premature contractions (class 2) is the least accurate arrhythmia when using the Adam optimizer. This results can be due to the computation of gradients using this optimiser, as so, the gradient may reach a local minima but not the lowest point. In this case, it may be because of the small number of beats of this class on the test dataset. On the other side, SGD is able to find better gradients for the classification of all the classes. However, it can be seen that when the model was trained with 100 epochs it was able to reach better results when classifying the atrial premature contractions (class 2).

After the best model was trained and evaluated, the new dataset, with unforeseen data for the model, was tested. These tests were performed record by record, where 21 records achieved more than 90% of accuracy. The only problem detected was when the model tried to predict class 2 where there was no class 2 in any of those records. Tacking this into account, a new model was trained, comprising the 11 least accurate records for training and the remaining 5 records with an accuracy below 30% for testing, we decide not to use all the records due to the lack of computational power. Table 2 presents the precision, recall, and F1-score values of the last evaluation of the new model.

This model was able to accurately classify the arrhythmias with an accuracy of 81% while the F1-scores for the normal class (class 0) reached 89.9%, the other two classes performed poorly, reminding the need to perform data augmentation to balance the data to train the model.

On the other hand, several limitations raised during the development of this project, such as, the lack of computational power to diminish the training time, which took over 15 hours for the subset of only 11 records from the second dataset as well as the imbalanced dataset. These issues present a hindrance in the success of these approaches, nevertheless it is important to emphasise it presents an advance in relation to typical systems were the annotations have to be carefully reviewed 100% of the time.

Table 2 – Results Obtained for the New Dataset Based on 11 Records from the Hospital for Training and the Five Worst Records from the Hospital for Testing

Labels	Precision	Recall	F1 Score
Class 0	98.9%	82.3%	89.8%
Class 1	8.8%	47%	14.9%
Class 3	4.1%	35.3%	7.4%

In comparison with the related work, namely the work performed by [16], where the authors performed a beat-to-beat classification using the MIT-BIH arrhythmia database and obtained an accuracy of 83.4% in classifying the signal as normal or abnormal (arrhythmic), we were able to outperform their approach in 10% while classifying into four different classes.

Conclusions

In this paper, we developed a deep learning model that was able to accurately classify the heartbeat into four different classes. Focusing on two types of arrhythmia, the results obtained for this classification task were promising showing that this path for beat classification should be further investigated and providing a basis for future studies. Researchers have been dealing with beat classification using deep learning, however, the results were slowly reaching 90% of accuracy and many of these results were based on two-dimensional layers. This paper presents a ResNet with one-dimensional convolutional layers and was able to reach over 90% of accuracy and F1-scores for the four classes proposed, without falling on the rabbit-hole of overfitting.

A new dataset was created, having records 48 times bigger than the records from the MIT-BIH Arrhythmia Database. This allowed us to further explore these results, increasing our training data to create a new and better model, able to classify these 24-hour record accurately and also to try to balance the data available. Since deep learning is a methodology data-driven, the larger the dataset, the better the results. With this in min, this dataset will allow us to increase the spectrum of classified types of arrhythmia in order to create a fully automatic system without neglecting the precision and the importance of the outcome.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project: UID/EEA/50014/2019.

References

[1] A.C. Guyton, and J.E. Hall, TextBook in Medical Physiology, Elsevier Saunders, 2006.

[2] G.J. Tortora, and B. Derrickson, Principles of Anatomy & Physiology 14th Edition, 2014.

[3] J. Pan, and W.J. Tompkins, A Real-Time QRS Detection Algorithm, *IEEE Trans. Biomed. Eng.* **BME-32** (1985) 230–236.

[4] K. R., I. E., K. A., and P. D., CardIO library for deep research of heart signals, (2017).

[5] Welcome to BioSPPy — BioSPPy 0.4.0 documentation, (n.d.). https://biosppy.readthedocs.io/en/stable/ (accessed February 1, 2018).

[6] S.H. Jambukia, V.K. Dabhi, and H.B. Prajapati, Classification of ECG signals using machine learning techniques: A survey, *2015 Int. Conf. Adv. Comput. Eng. Appl.* (2015) 714–721.

[7] C.K. Roopa, and B.S. Harish, A Survey on various Machine Learning Approaches for ECG Analysis, **163** (2017) 25–33.

[8] A.-B.M. Salem, K. Revett, and E.-S. El-Dahshan, Machine learning techniques in electrocardiogram diagnosis, *Int. Multiconference Comput. Sci. Inf. Technol.* (2009) 429–433.

[9] A. Batra, and V. Jawa, Classification of Arrhythmia Using Conjunction of Machine Learning Algorithms and ECG Diagnostic Criteria, *Int. J. Biol. Biomed.* **1** (2016) 1–7.

[10] V. Priyadharshini, and S. Saravana, An Enhanced Approach on ECG Data Analysis using Improvised Genetic Algorithm, (2015) 1248–1256.

[11] T. Barman, R. Ghongade, and A. Ratnaparkhi, Rough Set based Segmentation and Classification Model for ECG, (2016) 18–23.

[12] M. Zihlmann, D. Perekrestenko, and M. Tschannen, Convolutional Recurrent Neural Networks for Electrocardiogram Classification, (2017) 1–4.

[13] B. Pyakillya, N. Kazachenko, and N. Mikhailovsky, Deep Learning for ECG Classification, *J. Phys. Conf. Ser.* **913** (2017) 012004.

[14] P.A. Warrick, and M. Nabhan Homsi, Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection, *Physiol. Meas.* (2018) 0–20.

[15] M. Limam, F. Precioso, and U. Côte, Atrial Fibrillation Detection and ECG Classification based on Convolutional Recurrent Neural Network, *Comput. Cardiol. (2010).* **44** (2017) 1–4.

[16] S. G, S. K P, and V. R, Automated detection of cardiac arrhythmia using deep learning techniques, *Procedia Comput. Sci.* **132** (2018) 1192–1201.

[17] S. Savalia, and V. Emamian, Cardiac Arrhythmia Classification by Multi-Layer Perceptron and Convolution Neural Networks, (2018).

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, (2014) 1–9.

[19] G. Sannino, and G. De Pietro, A deep learning approach for ECG-based heartbeat classification for arrhythmia detection, *Futur. Gener. Comput. Syst.* **86** (2018) 446–455.

[20] M. Ayar, and S. Sabamoniri, An ECG-based feature selection and heartbeat classification model using a hybrid heuristic algorithm, *Informatics Med. Unlocked*. (2018) 1–9.

[21] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, C. Bourn, and A.Y. Ng, Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks, (2017).

[22] S. Kiranyaz, T. Ince, and M. Gabbouj, Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks, *IEEE Trans. Biomed. Eng.* **63** (2016) 664–675.

[23] K. Luo, J. Li, Z. Wang, and A. Cuschieri, Patient-Specific Deep Architectural Model for ECG Classification, *J. Healthc. Eng.* **2017** (2017).

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

[25] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, and H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals, *Circulation*. **101** (2000) e215–e220.

[26] Physionet, The WFDB Software Package, (n.d.). https://physionet.org/physiotools/wfdb.shtml (accessed April 2, 2018).

[27] R. Rodrıguez, A. Mexicano, J. Bila, S. Cervantes, and R. Ponce. Feature extraction of electrocardiogram signals by applying adaptive threshold and principal component analysis. *Journal of Applied Research and Technology*, 13(2):261–269, 2015.

[28] Xinqi Louis Wang and J Mikael Eklund Smieee. A real-time ECG feature detection algorithm. *Ieee*, pages 5–7, 2014.

[29] Abhinav Vishwa, Mohit K. Lal, Sharad Dixit, and Pritish Vardwaj. Clasification of Arrhythmic ECG Data Using Machine Learning Techniques. *International Journal of Interactive Multimedia and Artificial Intelligence*, 1(4):67, 2011.

[30] Ali Isin and Selen Ozdalili. Cardiac arrhythmia detection using deep learning. *Procedia Computer Science Procedia Computer Science Procedia Computer Science*, 120(00):268–275, 2017.

[31] A.V. Oppenheim, A.S. Willsky, and S.H. Nawab, Signals and Systems, Prentice Hall, 1997.

[32] S.H. Jambukia, V.K. Dabhi, and H.B. Prajapati, Classification of ECG signals using machine learning techniques: A survey, *2015 Int. Conf. Adv. Comput. Eng. Appl.* (2015) 714–721.

[33] T.J. Jun, H.J. Park, N.H. Minh, D. Kim, and Y.H. Kim, Premature ventricular contraction beat detection with deep neural networks, *Proc. - 2016 15th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2016*. (2017) 859–864.

[34] A. Vishwa, M.K. Lal, S. Dixit, and P. Vardwaj, Clasification Of Arrhythmic ECG Data Using Machine Learning Techniques, *Int. J. Interact. Multimed. Artif. Intell.* **1** (2011) 67.

[35] A. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, The marginal value of adaptive gradient methods in machine learning, *Advances in Neural Information Processing System.* (2017) 4148-4158.

Address for correspondence

Cláudia Brito: claudia.v.brito@inesctec.pt

António Sousa: als@di.uminho.pt