Person identification using heart rate and activity from consumer-grade wearables.

How do different types of cardiac diagnosis affect the accuracy of Deep Neural Networks to identify individuals by their heart rate?

Jip Kasper Harthoorn[1]

Supervisors: David Tax[1]**, Arman Naseri Jahfari**[1]**, Ramin Ghorbani**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Jip Kasper Harthoorn
Final project course: CSE3000 Research Project
Thesis committee: David Tax[1], Arman Naseri Jahfari[1], Ramin Ghorbani[1], Guohao Lan[1]

Abstract

Advancements in the precision and accuracy of consumer-grade wearables, such as a Fitbit, have enabled the identification and therefore authentication of individuals based on their emitted heart frequencies using these wrist-worn devices. With this type of authentication, a password is essentially sent out every second. This makes it a perfect form of authentication in fields where constant authentication is crucial. However, not much is known about how different types of cardiac diagnosis (e.g. fit or obese) influence the accuracy of this type of authentication. In this paper, it will be shown how different types of cardiac diagnosis affect the accuracy of Deep Neural Networks to identify individuals by heart rate. This study is done with data obtained from 14 subjects, having different types of cardiac diagnosis. A deep neural network consisting of multiple convolutional layers is being used to conduct the experiments. It has been shown that subjects with a paroxysmal atrial fibrillation diagnosis improve the accuracy the most, compared to the reference (normal healthy) subjects. On the other hand, (very) fit subjects decrease the accuracy the most. Heart failure and obese subjects have a similar accuracy compared to reference subjects.

1 Introduction

There are various ways to identify people using biometrics; identification based on physical characteristics, for example using a fingerprint. This is mainly used for authentication purposes. The problem with the current widely used biometrics systems is that people actively have to identify themselves. For example, with a fingerprint recognition system, one has to actively put a finger on the fingerprint scanner. Advancements in the precision and accuracy of consumer-grade wearables [1], such as a Fitbit, have enabled the identification and therefore authentication of individuals based on their emitted heart frequencies using these wrist-worn devices. Person identification using heart rate and activity from consumer-grade wearables can solve the problem of constant authentication; people are constantly being identified by their heart rate and/or step count. In other words, heartbeats offer continuous authentication. Whereas we enter passwords or scan our fingers once to access secure applications, a heartbeat could effectively send out a password every second. Another benefit, next to the one of constant authentication, is that everyone can make use of this type of identification, since the heartbeat is a bio-signal that is present in all human beings without exception. This can be useful in fields where only authorized people are allowed and perform certain tasks.

There have been various research done on this topic, also using consumer-grade wearables. A study by Nadas et al. [2] with two persons achieved an accuracy of 94% using a consumer-grade wearable. However, they concluded that more research is needed in order to provide more results with different datasets to consolidate the technology. Another well-known study [3] used an identification technique based on electrocardiograms (ECGs) and musical features (e.g., dynamics, rhythm or timbre). This study showed an accuracy of 96.6% using 18 subjects. This shows that identification based on heart rate is possible, but it does not show that it is accurate using a consumer-grade wearable.

In this paper, the following question will be researched: “How do different types of cardiac diagnosis affect the accuracy of deep neural networks to identify individuals by their heart rate?”. This is relevant because Deep Neural Networks are mostly used to identify individuals by heart rate. However, it is important to know if the accuracy will change depending on the cardiac diagnosis (e.g. obese). If these different types of diagnosis will have a noticeable effect on the accuracy of identifying individuals, then models should maybe be adapted to deal with this.

The rest of the paper is organized as follows: in Section 2 the methodology will be discussed. The experimental setup is presented in Section 3, describing more details about how the experiments were conducted. Section 4 contains the results, while Section 5 focuses on responsible research. Finally, conclusive remarks are drawn in Section 6.

2 Methodology

This section will include a general overview of how the research has been conducted. Subsection 2.1 will show the general approach, while subsection 2.2 shows the deep neural network architecture.

2.1 Approach

To measure the effect of different type of cardiac diagnosis on the accuracy, groups of subjects with different type of cardiac diagnoses have been selected. There are four types of special cardiac diagnosis:

- (Very) fit;
- Obese;
- Heart failure;
- Paroxysmal atrial fibrillation (irregular heart rate).

There also is a reference group. Therefore, there are five test groups in total.

Next up, the heart rates of the groups are being fed into the deep neural network (DNN); the groups will be fed into the DNN seperately. The objective of the DNN is to undergo training using heart rate information, enabling it to classify individuals based on new heart rate data once the training is complete. The testing accuracies of this classification task will be measured. To ensure reliability, the task will be repeated multiple times with new runs (initializations) of the DNN, and the test accuracy results will be averaged. After that, the new other groups will undergo the same approach whereafter the results will be plotted and analysed.

2.2 Deep Neural Network Architecture

To address the research objectives and effectively analyze the normalized data, a DNN was implemented. PyTorch, a popular Python library for deep learning, was utilized for building and training the DNN model.

Given the sequential nature of the input data, one might naturally think that recurrent neural networks (RNNs) would be well-suited for the task. However, it has been demonstrated that training RNNs can be challenging due to convergence issues [4]. Moreover, RNNs cannot be parallelized, resulting in slower inference times. In contrast, 1D convolutional layers can be fully parallelized and are capable of capturing local context within sequences. By stacking multiple convolutional layers, we can effectively capture a broader context, which leads to a fixed-sized representation of the entire sequence [5].

The proposed network (table 1 and figure 1) consiststed of two convolutional layers, followed by a pooling layer each with a kernel size of 2. The ReLU activation fuction was used for both layers. They also both had a dropout of 0.1. The first input layer had a single input channel (heart rate channel), four output channels, a kernel size of 5 and a stride of 1. The second convolutional layer had four input channels, eight output channels, a kernel size of 128 and a stride of 32. These layers are followed by a flatten layer, then followed by a LSTM layer, which in turn is followed by a fully connected layer. Lastly, the Softmax activation function is used, which is recommend to use in the output layer [6].

Layer Name	Hyperparameter settings
Conv1D	Kernel Size = 5, Stride = 1, Activation = ReLU
Dropout	Dropout Rate = 0.1
MaxPool1D	Pool Size = 2
Conv1D	Kernel Size = 128, Stride = 32, Activation = ReLU
Dropout	Dropout Rate = 0.1
MaxPool1D	Pool Size = 0.1
Flatten	-
LSTM	Output Channels = 128
Linear	Out Channels = 6, Activation = SoftMax

Table 1: Overview of the proposed network.

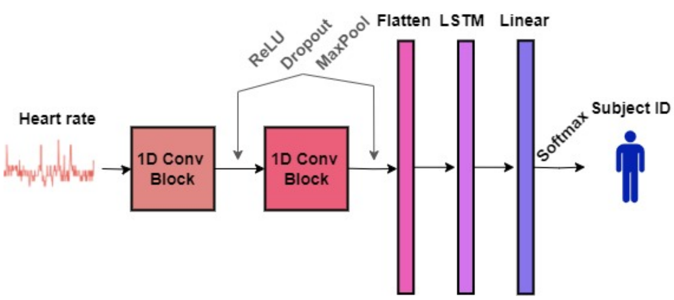

Figure 1: The proposed network; two convolutional layers followed by a flatten, LSTM and linear layer. The task is to classify a subject based on heart rate.

3 Experimental setup

This section describes the details of how the experiments or studies were conducted. The primary goal of this section is to provide enough information for other researchers to replicate the study and verify the results. Subsection 3.1 will describe how the data was collected. *More sections to come.*

3.1 Data collection

A diverse dataset from the Haga Ziekenhuis in The Hague, consisting of information from 14 subjects wearing different wearable devices (figure 2), was used for this research. The data containted heart rates and step counts, as well as meta data such as cardio status, gender, age. Heart rate has been resampled to 0.2 Hz (once every 5 seconds). The original sample rate was variable but 0.2 Hz was the most prevalent sample rate in the data set. The step counter was sampled once every minute.

Figure 2: Wearables used to obtain the heart rates and step counts; Fitbit Inspire 2, Fitbit Charge 5 and Fitbit Charge 2 respectively. *Source: www.fitbit.com*

3.2 Subject composition

To measure the effect of different type of cardiac diagnosis on the accuracy, groups of six subjects have been selected (figure 3). Two of those six subjects are subjects with a special type of cardiac diagnosis, the rest are so called 'reference' subjects; these subjects have no special cardiac diagnosis and can be seen as a baseline. The reason why there are two cardiac diagnosis subjects extended with four reference subjects is because there has only been enough data of two subjects with a special cardiac diagnosis. Next to that, if measuring the accuracy only between those two subjects (binary), the accuracy almost always converged to 100%. That is why it is extended with four reference subjects; to see the effect of the special subjects on the reference group. There has also been tests done with only reference subjects as a baseline. Therefore, there are five test groups in total.

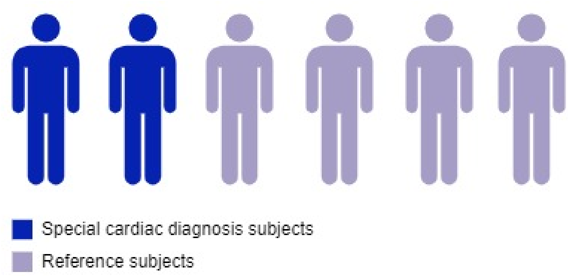

Figure 3: Subjects composition; two subjects with a special cardiac diagnosis and four reference subjects.

3.3 Window Size

In order to gather comprehensive data on an individual's heart rate pattern, a window size of one day was utilized, as depicted in Figure 4. This particular window size was chosen to ensure that it encompasses all the relevant activities that can affect a person's heart rate, ranging from daily walking routines to periods of sleep.

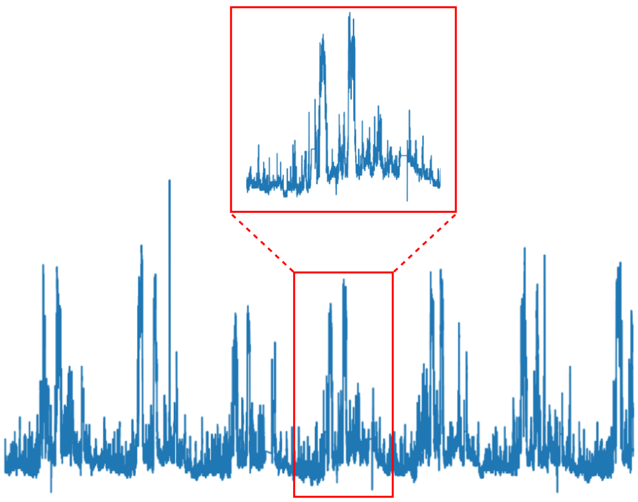

Figure 4: Raw heart rate data of a single subject; a window of 1 day is being taken.

3.4 Data preprocessing

To ensure reliable and comparable results, the collected data underwent necessary preprocessing steps. Heart rates were normalized. This had a significant impact on the testing accuracy (around 35% higher with normalization). The heart rate was normalized using mean and standard deviation (standardization).

To maintain consistency and prevent overfitting, the number of data points were constricted to 1,000,000. This was chosen because the subjects with the least data points had around 1,000,000 data points. This number comes down to around 50 days of data.

No feature selection has been used.

3.5 Hyperparameters

The input size is 14,000, this is around a day of data. A learning rate of 0.001 together with a batch size of 256 was chosen.

There is a high correlation between the learning rate and the batch size. When the learning rates are high (0.001), the large batch size (256) performs better than with small learning rates (0.0001). Based on a study of the effect of batch size on the generalizability of the convolutional neural networks [7], it is recommended to either take a small batch size with a small learning rate or a high batch size with a high learning rate. The latter gave the best results.

The number of epochs is 50, since the model converged just before that.

CrossEntropyLoss has been used as loss function together with Adam as optimizer, both with their default parameters.

The experiments were conducted using CUDA on a NVIDIA GeForce GTX 1660 Ti with Max-Q Design.

To reproduce and give more insight in the experiments, the code has been made open source and is available on GitHub[1].

4 Results

In this section, the results will be presented and reasoned about. Subsection 4.1 will show the general results of using the trained DNN on the 14 subjects. Next up, the cardiac diagnoses groups will individually be discussed. Subsection 4.2 will cover the paroxysmal atrial fibrillation group, 4.3 the obese group, 4.4 the (very) fit group and 4.5 the hear failure group.

4.1 General Results

Since the reference group has six subjects (table 1) and four of them are being added each run to the special cardiac diagnoses group, there are multiple combinations of reference subjects to add to the other groups.

Subject groups	**Subject IDs**
Reference	1, 2, 3, 4, 5, 6
(Very) fit	7, 8
Obese	9, 10
Heart failure	11, 12
Paroxysmal atrial fibrillation	13, 14

Table 2: Test

There are 15 combinations possible to make from 1, 2, 3, 4, 5, 6, when taking four of these each time:

- 1, 2, 3, 4; 1, 2, 3, 5;
- 1, 2, 3, 6; 1, 2, 4, 5;
- 1, 2, 4, 6; 1, 2, 5, 6;
- 1, 3, 4, 5; 1, 3, 4, 6;
- 1, 3, 5, 6; 1, 4, 5, 6;
- 2, 3, 4, 5; 2, 3, 4, 6;
- 2, 3, 5, 6; 2, 4, 5, 6;
- 3, 4, 5, 6.

To make the experiments more reliable, the average of each special cardiac diagnoses group has been taken from 45 runs (initializations of the DNN); each of the 15 combinations three times. The reference group has also ran 45 times without any extension of subjects. The results of these runs can be seen in figure 5.

The results show that the subjects with a paroxysmal atrial fibrillation diagnoses affect the training of the DNN in such a way that the test accuracy increases. On the other hand, the '(very) fit' group has an average ending accuracy of around 10% lower than the 'paroxysmal atrial fibrillation' group.

[1]https://github.com/Jip1912/dnn_hr_cardiac_diagnosis_analysis/blob/master/analysis.ipynb

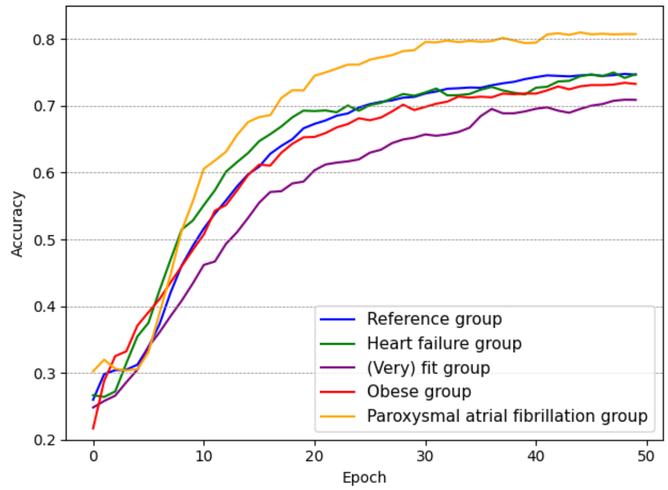

Figure 5: The average testing accuracies of 45 runs per subject group.

The 'heart failure', 'reference' and 'obese' group have a very similar average ending accuracy. Besides this, they all converge at around the same amount of epochs.

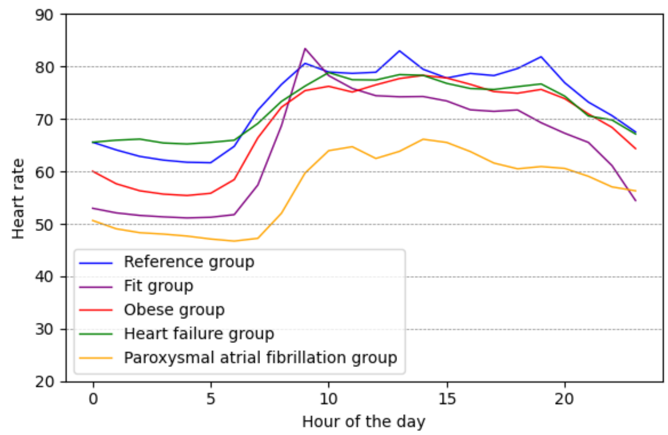

Figure 6: The average heart rates per hour per subject group.

4.2 Paroxysmal atrial fibrillation group

The question that arises is why the paroxysmal atrial fibrillation group has an average ending accuracy of around 10% higher than the (very) fit group. When looking at the average heart rates per hour (figure 6), it is clear that the paroxysmal atrial fibrillation group is an outlier.

The average heart rates of the paroxysmal atrial fibrillation group, ranges in a lower zone than the other groups. This might imply that the DNN has an easier time learning the paroxysmal atrial fibrillation group, since it is more unique than the rest. Based on data collected from a wrist-based wearable device, a study [8] has confirmed that atrial fibrillation induces distinct heart rate patterns. These patterns can be utilized by algorithms to detect and diagnose atrial fibrillation, highlighting their unique nature.

4.3 Obese group

When looking at the obese group, the results are quite similar to the reference group. Most obese individuals without any clinical heart disease have a normal heart rate, similar to individuals who are not obese [9]. However, as obesity levels increase, there is a slight rise in heart rate. It's important to note that tachycardia, which is an abnormally high heart rate, is not commonly observed in obese individuals without heart disease. A higher mean of the heart rate is not present at our subjects with obese. The findings from Stuart Frank et al., which show that most obese individuals have a normal heart rate, highlight the fact that the heart rate patterns in the obese group are similar to those in the reference group (non-obese individuals). This suggests that the accuracy of identifying obese individuals by their hear rate is comparable to that of individuals without obesity.

4.4 (Very) fit group

(Very) fit individuals have a similar heart rate pattern to individuals with a normal heart rate. However, their resting heart rate is often lower [10]. The latter can be clearly seen in figure 6; when looking at general sleeping times (00:00 - 06:00), the average heart rate of (very) fit people is significantly lower than the reference group. Regular aerobic exercise enhances autonomic nervous system balance, resulting in increased heart rate variability and improved stability of heart rate response [11]. This might imply that the heart rates of the (very) fit subjects are more stable and less diverse, which makes it harder for the DNN to distinguish this group and therefore results in a lower average testing accuracy.

4.5 Heart failure group

Neural networks have shown reliable detection of heart failure [12]. This suggests that heart failure has distinct patterns that differ from individuals with a normal heart rate. In other words, heart failure can be characterized by specific features that are distinguishable from those seen in individuals without heart failure, which is a similar scenario as with the paroxysmal atrial fibrillation group. However, the mean of the heart rate per hour of the heart failure group is the most similar to the reference group compared to other groups. This can explain the fact that the accuracy of identifying individuals with heart failure is similar to the reference group.

5 Responsible Research

In this section, the focused will be shifted to the important concept of responsible research. Working with person-related data can have ethical implications. Especially when this data can be used to identify them. Good and trustworthy research is built upon five essential principles from the Netherlands Code of Conduct for Research Integrity: honesty, thoroughness, transparency, independence, and accountability [13]. These principles provide the foundation for reliable scientific work and promote a trustworthy research community. The ethical implications, potential biases and reproducibility of the research will be discussed.

5.1 Ethical implications

Ethical concerns arise when using biometric data for identification purposes. Heart rate and step count data can reveal significant information about an individual [14]. Researchers and businesses should ensure that individuals provide informed consent before their heart rate and step count data are collected and used for identification.

Large business like Fitbit or even Google or Facebook could obtain biometric data and use this technology to authenticate people or get to know more about someone. In July of 2022, a survey was conducted and included about 4,000 respondents divided between the United States and several countries in Europe [15]. The results showed that more and more people are becoming concerned about their online privacy, with 60% of consumers expressing worries. Additionally, 53% feel like they have no control over their online identity. About 35% are tired of their personal data being used for targeted ads, and 25% find it creepy. These numbers reflect the increasing awareness and discomfort surrounding the use of personal information online. It is therefore important to prioritize the implementation of strong safeguards to protect individuals' online privacy, particularly regarding the collection and use of biometric data by major companies.

5.2 Potential bias

When training a DNN, it is important to have a diverse training group to prevent bias. If the training dataset exclusively consists of individuals who are very fit, it may disadvantage individuals who are obese or have different physical characteristics. In this research, a dataset has been used with varying subjects to ensure that there is no bias on a specific charasteristic of a person. Another type of bias could be caused by the type of wearable that has been used to collect the data. A wearable device can have an unique kind of noise which can cause the model to learn the device rather the person. This has been a prevalent problem in an earlier mentioned study: "One major hindrance towards our goal was the existence of unique sensor noise, which misled the DNN to classify the sensor instead of the person" [5]. As mentioned earlier, the data has been obtained from 3 different type of devices, which presumably lowers this risk.

6 Conclusions and future work

In this paper, the following question has been discussed: "How do different types of cardiac diagnosis affect the accuracy of deep neural networks to identify individuals by their heart rate?". Four types of cardiac diagnoses together with one reference group have been used to conduct multiple experiments, from which the results have been averaged.

Based on the results described in section 4, using 14 subjects, the subjects with the 'paroxysmal atrial fibrillation' diagnoses affect the accuracy in such a way that the average was the highest. Thereafter, the 'heart failure' and reference group had the highest accuracy, followed by the 'obese' and '(very) fit' group respectively.

The hypothesis is that the paroxysmal atrial fibrillation has the highest accuracy, because subjects suffering from paroxysmal atrial fibrillation have an irregular and more unique heart rate pattern compared to the other subjects, which means that the DNN has an easier time to recognize these subjects. More research is needed to confirm this.

It is crucial for future studies to include a larger number of individuals with special cardiac diagnoses while conducting experiments. In this research, each group contained only two subjects with a special cardiac diagnoses, which can make the results less reliable.

References

[1] Fuller D. et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: Systematic review. In *JMIR Mhealth Uhealth*, volume 8, page e18694, 2020.

[2] Joäo Nadas et al. Towards continuous subject identification using wearable devices and deep cnns. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, 2020.

[3] Carmen C. et al. Ecgsound for human identification. *Biomedical Signal Processing and Control*, 72:103335, 2022.

[4] Razvan Pascanu et al. On the difficulty of training recurrent neural networks. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[5] G. Retsinas et al. Person identification using deep convolutional neural networks on short-term signals from wearable sensors. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3657–3661, 2020.

[6] Tomasz Szandała. *Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks*, pages 203–224. Springer Singapore, Singapore, 2021.

[7] Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, 6(4):312–315, 2020.

[8] A. G. Bonomi et al. Atrial fibrillation detection using a novel cardiac ambulatory monitor based on photoplethysmography at the wrist. *Journal of the American Heart Association*, 2018.

[9] Stuart Frank et al. The electrocardiogram in obesity: Statistical analysis of 1,029 patients. *Journal of the American College of Cardiology*, 7(2):295–299, 1986.

[10] Anne Kerstin Reimers et al. Effects of exercise on the resting heart rate: A systematic review and meta-analysis of interventional studies. *Journal of Clinical Medicine*, 7(12), 2018.

[11] Arto Hautala, Antti Kiviniemi, and Mikko Tulppo. Individual responses to aerobic exercise: The role of the autonomic nervous system. *Neuroscience and biobehavioral reviews*, 33:107–15, 05 2008.

[12] Jad Botros et al. Cnn and svm-based models for the detection of heart failure using electrocardiogram signals. *Sensors*, 22(23), 2022.

[13] ALLEA. The european code of conduct for research integrity. *Berlin-Brandenburg Academy of Sciences and Humanities*, 2017.

[14] Z. Xu et al. Evaluation of physical health status beyond daily step count using a wearable activity sensor. In *npj Digital Medicine*, volume 5, page 164, 2022.

[15] Datagrail. Consumer privacy in the age of personalization, privacy rights, and personal data. *The Great Privacy Awakening*, July 2022.