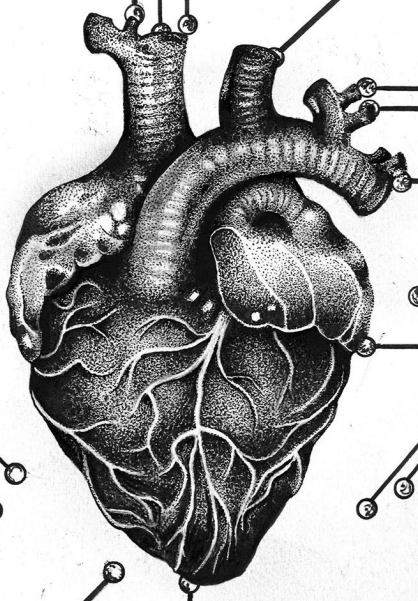
Improving Remote Cardiovascular Care with Wearables

Algorithms, Study Design, and Subject-Specific Adaptation



Arman Naseri Jahfari

Improving Remote Cardiovascular Care with Wearable Data

Algorithms, Study Design, and Subject-Specific Adaptation

Improving Remote Cardiovascular Care with Wearable Data

Algorithms, Study Design, and Subject-Specific Adaptation
Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Tuesday 13 January 2026 at 10:00 o'clock

by

Arman NASERI JAHFARI

Master of Science in Electrical Engineering, Delft University of Technology, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, *promotor*
Dr. D.M.J. Tax,	Delft University of Technology, *copromotor*
Dr. I.A.C. van der Bilt,	University Medical Center Utrecht, Haga Teaching Hospital, Den Haag,*copromotor*
Independent members:	
Prof. dr. J.J. van den Dobbelsteen	Delft University of Technology
Prof. dr. ir. M.S. Kleinsmann	Delft University of Technology
Dr. ir. M. Mulder	Erasmus University Medical Center, Rotterdam
Prof. dr. P. van der Harst	University Medical Center Utrecht
Dr. J. Urbano Merino	Delft University of Technology, reserve member

Keywords: Wearables, remote cardiovascular monitoring, machine learning
Printed by: Proefschriftspecialist
Cover by: Karim Bahri

ISBN 978-94-93483-60-6

An electronic copy of this dissertation is available at
https://repository.tudelft.nl/.

De tijd tikt langzaam door [...] Want enkel als ik slaap, ben ik even weg van al m’n dromen. Maar ik heb slapeloze nachten.

The Opposites

CONTENTS

SUMMARY

Cardiovascular diseases remain the leading cause of death worldwide, yet early detection and continuous monitoring remain challenging outside clinical settings. This dissertation is motivated by the growing potential of remote health monitoring to address this gap—specifically, the use of consumer-grade smartwatches to track cardiovascular health through physiological signals. Although consumer-grade wearables are traditionally merely used as fitness-oriented or recreational, this work investigates the clinical applicability of smartwatch-derived signals for disease monitoring in real-world, non-clinical environments. By enabling scalable, data-driven detection of cardiovascular conditions in everyday settings, such a system has the potential to reduce the burden on physicians, provide patients with continuous insights, and alleviate pressure on healthcare systems through earlier intervention and more personalized care.

By assessing how far wearable-based research has progressed toward operational deployment and identify critical shortcomings in real-world utility and generalizability, we confront several major challenges intrinsic to this domain: the medical interpretability of noisy consumer-grade signals, high inter-subject variability, and the inherent complexity of time-series data that varies with context (e.g., day/night cycles, physical activity).

Our solution strategy is grounded in machine learning techniques that aim to learn robust, transferable representations of physiological data. In particular, we explore contrastive learning, weak supervision, and morphological modeling—such as acceleration-deceleration curve analysis—as tools to extract clinically relevant patterns. These methods are evaluated across both publicly available and proprietary datasets to ensure applicability to diverse populations.

By addressing these challenges, this dissertation advances the case for smartwatches as viable tools for longitudinal, data-efficient cardiovascular monitoring, contributing to a future in which early detection of conditions like atrial fibrillation and heart failure is feasible at scale in everyday settings.

SAMENVATTING

Hart- en vaatziekten blijven wereldwijd de belangrijkste doodsoorzaak, terwijl vroege opsporing en continue monitoring buiten de klinische setting nog steeds een uitdaging vormen. Dit proefschrift is geïnspireerd door de toenemende mogelijkheden van gezondheidsmonitoring op afstand om hierin verbetering te brengen—specifiek door het gebruik van consumentgerichte smartwatches om de cardiovasculaire gezondheid te volgen via fysiologische signalen. Hoewel wearables voor consumenten traditioneel vooral worden gebruikt voor sportieve of recreatieve doeleinden, onderzoekt dit werk de klinische toepasbaarheid van smartwatchsignalen voor het monitoren van ziekten in realistische, niet-klinische omgevingen. Door schaalbare, datagestuurde detectie van cardiovasculaire aandoeningen in het dagelijks leven mogelijk te maken, kan een dergelijk systeem de belasting voor artsen verminderen, patiënten continu inzicht bieden, en de druk op zorgsystemen verlichten via vroegtijdige interventie en meer gepersonaliseerde zorg.

Door te analyseren hoe ver onderzoek naar wearables is gevorderd richting daadwerkelijke toepassing en welke belangrijke tekortkomingen er nog bestaan op het gebied van bruikbaarheid en generaliseerbaarheid in de praktijk, pakken we verschillende kernuitdagingen aan die eigen zijn aan dit domein: de medische interpreteerbaarheid van ruisgevoelige signalen van consumentenapparaten, grote interindividuele variatie, en de inherente complexiteit van tijdreeksdata die afhankelijk is van context (bijv. dag-/nachtcycli, fysieke activiteit).

Onze oplossingsstrategie is gebaseerd op machine learning-technieken die gericht zijn op het leren van robuuste en overdraagbare representaties van fysiologische gegevens. We onderzoeken in het bijzonder contrasterend leren, zwakke supervisie en morfologische modellering—zoals de analyse van acceleratie-deceleratiecurves—als hulpmiddelen om klinisch relevante patronen uit de data te halen. Deze methoden worden getest op zowel openbare als eigen datasets om de toepasbaarheid op diverse populaties te waarborgen.

Door deze uitdagingen aan te pakken, versterkt dit proefschrift de positie van smartwatches als bruikbare hulpmiddelen voor langdurige, data-efficiënte monitoring van de cardiovasculaire gezondheid, en draagt het bij aan een toekomst waarin vroege opsporing van aandoeningen zoals atriumfibrilleren en hartfalen op grote schaal mogelijk is in het dagelijks leven.

1

INTRODUCTION

1

The study of the pulse, known as pulsology, has been a fundamental component of medical diagnosis since antiquity [1]. In ancient medical traditions across regions such as Egypt, Greece, and Persia, the pulse was recognized as a non-invasive indicator of internal physiological states and was systematically examined to assess health and detect disease.

One of the ancient pioneers was Ibn Sina and if he were our contemporary, he would likely have made significant contributions to heart rate analysis. He shaped the modern approach to examining the pulse using the wrist. He considered several parameters, such as strength and slowness, to devise different pulse profiles that describe the physiological condition of a person—for example, to identify cardiovascular disease [2–5]. Through this handcrafted pattern recognition, pioneers like Sina (and many others) paved the way for automated approaches to heart rate analysis.

In modern times, smartwatches enable automated, wrist-based, continuous and non-invasive monitoring of physiological signals, opening new avenues for health assessment and disease prevention. The key physiological parameters tracked by modern smartwatches are heart rate[6] and number of steps taken[7].

The subtle fluctuations in heart rate over time—known as heart rate variability (HRV)— carry a wealth of physiological information. HRV has emerged as a valuable second-order biomarker[8][9][10], reflecting dynamic fluctuations in sympathetic and parasympathetic nervous system activity, and has been linked to a wide range of physiological and pathological states[11][12]. In particular, HRV has shown clinical potential in the detection of cardiac conditions such as atrial fibrillation (AF) and heart failure (HF).

AF is a cardiac arrhythmia characterized by chaotic and rapid electrical activity in the atria, leading to irregular conduction to the ventricles[13]. As a result, the heart rhythm appears disordered, often accompanied by a highly elevated heart rate and by fluctuating intervals that lack coherent temporal structure. Consequently, the heart rate variability becomes highly erratic and unpredictable [14][15].

Heart failure is a condition in which the heart is unable to pump blood effectively and may result from weakened cardiac muscle[16]. The heart's ability to respond to physiological cues diminishes rather than the healthy fluctuation that indicates a responsive cardiovascular system, the heart rhythm in HF often becomes rigid and monotonous, indicating a reduced heart rate variability[17][18].

1.1. MEASURING HEART RATE: FROM HEART TO WRIST

Traditionally, the heart rate is estimated from the electrocardiogram (ECG) and some smartwatches also feature ECG measuring functionality,

allowing users to record short-term ECG signals by placing their fingers on the device's electrodes, while remaining still. ECG measures the electrical activity of the heart and an ECG cardiac cycle is characterized by several distinct phases that reflect the electrical activity of the heart. Among these, the QRS complex is the most critical, as it represents ventricular depolarization, the process that triggers the contraction of the ventricles, enabling blood to be pumped to the lungs and the rest of the body.

Figure 1.1 illustrates the different phases of an ECG. The Q wave

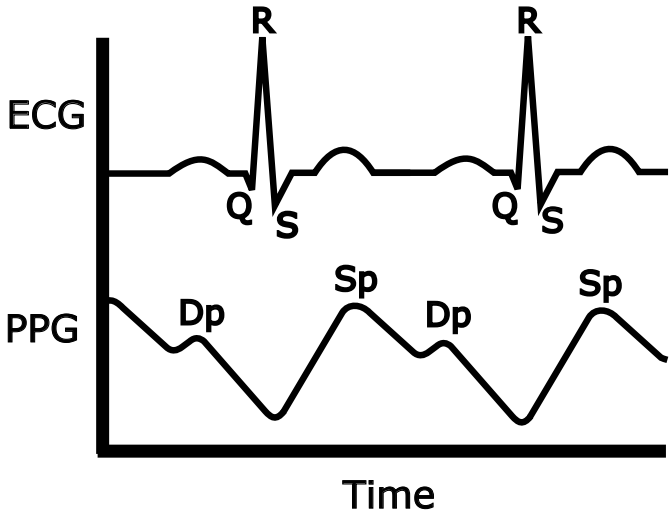


Figure 1.1.: The top panel shows an electrocardiogram (ECG) signal with Q, R, and S waves annotated. These correspond to different stages of ventricular depolarization: the Q wave marks initial depolarization, the R wave represents the peak of electrical activation, and the S wave indicates final depolarization of the ventricles. The bottom panel displays a photoplethysmogram (PPG) signal with labeled systolic peak (Sp) and diastolic peak (Dp), representing the primary and secondary peaks in peripheral blood volume. The R peak triggers the ventricular contraction that leads to the systolic peak in the PPG. Heart rate can be estimated from peak-to-peak time intervals; either R-to-R in the ECG or Sp-to-Sp in the PPG.

occurs due to the initial depolarization of the interventricular septum (the wall separating the left and right ventricles). As the electrical impulse spreads through this region, it generates a brief downward, negative signal before full ventricular activation.

The R wave represents the rapid depolarization of the ventricular myocardium (the main ventricular muscle). At this stage, a large portion of the ventricular muscle becomes electrically activated and as a result, this phase generates the most significant electrical change detected by the ECG, producing a sharp upward signal and the prominent R peak.

Finally, the S wave is the downward signal following the R wave, indicating the final phase of ventricular depolarization. This occurs as the electrical impulse spreads to the outermost regions of the ventricles and completes the activation of all muscle fibers.

Since the R peak within the QRS complex represents the strongest electrical signal generated by the heart, the R-R interval (the time between consecutive R peaks) is used to determine the heart rate. Because ECG measures cardiac electrical activity, it provides a very accurate assessment of heart rate and cardiac function.

Despite its accuracy, ECG measurement in smartwatches is often limited to manual activation while sitting still and short-duration recordings, making continuous monitoring impractical. The photoplethysmograph (PPG), on the other hand, allows for continuous, cost-effective and real-time tracking of pulse rate with minimal user intervention. This trade-off explains why most consumer-grade smartwatches primarily rely on PPG for heart rate estimation, while ECG functionality is typically reserved for on-demand assessments.

PPG is an optical technique that detects blood volume changes in the microvascular tissue of the skin. This method involves emitting light—typically green or infrared—into the skin and measuring the amount of light that is either absorbed or reflected back by the blood flow. Since blood absorbs more light than surrounding tissue, fluctuations in light absorption correspond to the pulsatile changes in blood volume driven by the cardiac cycle[19].

The ventricular depolarization (measured as R peak in the ECG) triggers a powerful contraction that forces blood out of the heart into the pulmonary artery and aorta. This ejection creates a pressure wave that travels rapidly through the arterial walls (faster than the blood itself moves). When the pressure wave reaches the wrist, it causes the arterial walls to expand elastically. This expansion creates a slight increase in local blood volume: the stretched space draws in a small amount of blood from adjacent regions. PPG sensors detect this volume change optically: more blood in the area means more light is absorbed, less is reflected, and this is recorded as the systolic peak in the PPG waveform.[20][21]

Conversely, during the diastolic phase, as the heart relaxes and arterial blood volume decreases, light absorption diminishes, which allows more light to be reflected or transmitted back to the detector. This phase is associated with the descending edge of the PPG waveform, leading to a trough, which represents the point of minimum blood

volume (diastolic peak) before the next systolic upstroke[19].

Note in Figure 1.1 that there is a delay between the R peak and the systolic peak (pulse transit time), caused by the time it takes for the heart to contract after depolarization and the resulting pressure wave to reach the wrist.

Whether heart rate is measured via R-R intervals in ECG or systolic peak-to-peak (Sp-Sp) intervals in PPG, both approaches capture the time between successive heartbeats, which we will collectively refer to as the inter-beat interval (IBI).

1.2. Diversity of heart rate variability quantification methods

The extraction of heart rate from the IBI enables the derivation of HRV, which has been shown to be predictive of cardiovascular disease in both clinical and wearable contexts. Numerous short-time methods have been proposed for quantifying HRV and are categorized by the Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (TFESCNASP)[22][23], outlined in 1.1, each capturing different aspects or perspectives of the underlying variability.

Table 1.1.: Common methods for computing Heart Rate Variability (HRV), categorized by type. N is the total number of inter-beat intervals (IBI) considered (either derived from ECG or PPG) and N_b is the number of beats in bin b. HR is heart rate. ϕ represents the average pairwise similarity computed over a set of inter-beat intervals (IBIs); spectral methods are likewise applied to this set.

Method	**Formula**	**Type of Method**
RMSSD	$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(IBI_{i+1}-IBI_i)^2}$	Statistical
SDNN	$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(IBI_i-\overline{IBI})^2}$	Statistical
pNN50	$\frac{\#(\lvert IBI_{i+1}-IBI_i\rvert>50\text{ms})}{N}$	Statistical
HRV Triangular Index	$\frac{N}{max(N_b)}$	Geometric
Poincaré Plot SD1	Diagonal Std. dev.	Geometric
Poincaré Plot SD2	Off-diagonal Std. dev.	Geometric
LF power	Spectral power in 0.04-0.15 Hz band	Spectral
HF power	Spectral power in 0.15-0.4 Hz band	Spectral
LF/HF Ratio	$\frac{\text{LF power}}{\text{HF power}}$	Spectral
MF (Mid Freq)	Power in ∼0.07–0.14 Hz band	Spectral
DFA	Residual variance after detrending	Non-linear
Approximate Entropy	$\phi_{m+1}(r)-\phi_m(r)$	Non-linear
Valsalva ratio	$\frac{HR_{max}}{HR_{min}}$	Morphological

Statistical methods rely on (several) time differences between successive beats. The Root Mean Square of Successive Differences (RMSSD)[24] captures the change in heart rate between two subsequent beats and this is averaged for N successive interval pairs. The Standard Deviation of NN intervals (SDNN) is similar, but is computed relative to $\overline{RR}$, the average RR interval over all N beats. Finally, pNN50 measure the proportion of RR intervals in which the change in successive RR interval is greater than 50 ms[25].

Geometrical methods rely on approximating distributions and quanti-

fying patterns from them. The HRV Triangular Index[26] considers the histogram of the duration of beats (i.e. time between adjacent R peaks) and is defined as the ratio of the total number of beats w.r.t. the number of beats in the modal bin. A Poincaré plot[27] is a two-dimensional representation in which each RR interval is mapped against the subsequent one, i.e., $(IBI_n, IBI_{n+1})\ \forall\, n \in [0, N-1]$, characterizing beat-to-beat heart rate dynamics. The representation is typically ellipsoidal and to quantify this, the standard deviation is calculated from two different axes. SD1 is the standard deviation where the data is projected on the anti-diagonal ($IBI_{n+1} = -IBI_n$) and captures short-term variability. Similarly, SD2 is the standard deviation where the data is projected on the diagonal ($IBI_{n+1} = IBI_n$) and captures long-term variability. The type of variabilities captured can be intuitively grasped by considering the case where most data points are very close to the diagonal, but spread over many values of IBI_n. In this case SD2 is large (many different values for the heart rate occur), but SD1 is small (the successive heart rate does not deviate much from the previous heart rate). In contrast, when the data density approximates a circle with small spread along the diagonal direction, but a relatively large spread along the anti-diagonal, this implies that the heart rate fluctuates heavily around the same heart rate value.

Spectral methods decompose the time series of successive IBIs into frequency components. Two important frequency bands[22] are recommended by the TFESCNASP for the purpose of spectral analysis of heart rate variability. The low frequency (LF) component (ranging from 0.04 to 0.15 Hz) reflects a mix of signals from both the sympathetic and parasympathetic nervous systems, while the HF component (0.15 to 0.40 Hz) is mainly linked to parasympathetic activity, especially the influence of the vagus nerve[28][29]. The LF/HF ratio is commonly used as an indicator of the balance between these two branches of the autonomic nervous system. A mid-frequency (MF) band (ranging from 0.05 to 0.20 Hz) is used to analyze specific oscillatory patterns such as Mayer waves[30]. The spectral decomposition has been done using Short-time Fourier Transforms[31], autoregressive models[32][33] and Discrete Wavelet Transforms[34].

The TFESCNASP further defines methods that describe the regularity or complexity of time series non-linearly. Approximate entropy[35] essentially assesses average pairwise similarity (ϕ) between segments of IBIs of size m, where two segments are considered similar if they differ up to an amount defined by r. The similarity ϕ is quantified using the Chebyshev norm (also known as the L^∞ norm) which takes the maximum absolute difference between corresponding elements of the two segments. The difference between $m+1$ and m is considered to describe how much the similarity changes when a longer segment size is chosen. Drastic similarity changes are interpreted as 'unpredictable'

and thus have high approximate entropy. Sample entropy is almost identical, but does not include the similarity of a segment with itself. Detrended Fluctuation Analysis (DFA) characterizes long-range correlations in non-stationary timeseries[36], useful for distinguishing between healthy and pathological dynamics.

Finally, the Valsalva maneuver[37] provokes an autonomic nervous response through breathing. The Valsalva ratio is calculated as the maximum heart rate during the maneuver, divided by the minimum heart rate observed within 30 seconds following that peak. Similarly, acceleration and deceleration dynamics of heart rate[38] are autonomic responses to e.g. physical activity. Since this requires knowledge about the shape of the resulting heart rate curve, reminiscent of ECG morphology, we denote these type of HRV measures as morphology methods. Morphological methods are particularly well-suited for smartwatch data, as they rely on broader (temporal) patterns and are therefore less sensitive to the lower accuracy and slower sampling rates of consumer-grade devices.

1.3. CHALLENGES & OPPORTUNITIES OF SMARTWATCH HEART MONITORING

The continuous monitoring capabilities of smartwatches help address several challenges in detecting heart disease. In its early stages, heart rhythm disorders can be episodic, with spontaneous onset and termination, making it difficult to detect through sporadic measurements. Continuous heart rhythm monitoring increases the likelihood of capturing such transient events. More broadly, smartwatches hold significant potential as early-warning or even prognostic tools for cardiovascular health. By enabling the long-term observation of subtle physiological changes, they offer a means to detect early signs of deterioration and may help mitigate the progression of serious heart disease altogether. PPG does not measure heart rate directly but instead captures pulse rate. Since pulse rate is derived from the blood flow dynamics at the wrist rather than directly from cardiac electrical activity, it may be influenced by vascular properties and motion artifacts, making it different from the heart rate measured via ECG and may reflect the underlying heart condition differently.

Furthermore, consumer-grade devices are not capable of continuously capturing and transmitting raw PPG waveforms due to constraints related to battery life, storage, and data bandwidth. Instead, heart rate is typically derived on-device from the PPG waveform by identifying the systolic peaks through signal processing. The heart rate can then be derived, based on the time difference between successive peaks, the inter-beat interval (IBI). Only the estimated heart rate values are

transmitted for external analysis. This presents a major limitation: while raw ECG waveforms contain rich information about the different phases of the cardiac cycle—information that is already partially lost in PPG—further reducing the signal to a single heart rate value removes critical diagnostic content altogether. Therefore, leveraging HRV is essential for extracting meaningful insights about cardiac health from the simple heart rate signal provided by a smartwatch.

Using HRV to detect AF and HF presents significant challenges due to the inherent variability in physiological and contextual factors across individuals and moments in time. HRV is influenced not only by underlying cardiac function but also by age, sex, body mass index (BMI), circadian rhythms, emotional state, physical activity, and even breathing patterns. This makes it difficult to define universal thresholds for 'normal' or 'abnormal' HRV across a diverse population. However, this complexity is precisely where machine learning algorithms excel: they can consider multiple HRV measures simultaneously and determine a complex threshold within the resulting high-dimensional space. Thus, in order to utilize smartwatch technology for cardiovascular outcomes to the fullest, we need a data-driven approach using machine learning.

In order to train a machine learning model, examples must be provided in the form of heart rate sequences paired with corresponding cardiovascular outcomes as labels. However, labeling longitudinal smartwatch data poses practical and clinical challenges. It would require physicians to review large volumes of time series data, which are often not sufficiently informative to make a diagnosis. As a result, smartwatch datasets are typically sparsely labeled with only a single diagnosis-level label per subject (e.g., atrial fibrillation, heart failure, or no known condition). Machine learning models must therefore be capable of learning from such weak or coarse supervision, where the presence or absence of a cardiovascular condition is known, but its precise manifestation over time is not labeled.

1.4. CONTRIBUTION OF THE THESIS

This thesis presents a series of algorithms and analyses designed to leverage HRV for the classification of various cardiovascular conditions, including heart failure and atrial fibrillation. By advancing automated interpretation of physiological signals captured via consumer-grade wearables, we aim to improve remote cardiovascular monitoring and contribute to the broader domains of wearable health technologies and personalized medicine. Chapter 1 offers a systematic review of the literature on cardiovascular machine learning applications in the context of wearable devices, identifying key trends, challenges, and gaps to deploy wearables, and in particular smartwatches, in healthcare. This is done by adapting the Technology Readiness Level (TRL) framework in

the context of wearables and by assessing included studies accordingly. Chapter 2 introduces the ME-TIME study, a prospective, longitudinal cohort study designed to collect continuous, real-world data from patients with heart failure and atrial fibrillation in free-living conditions, using commercially available Fitbit devices. It details the data acquisition protocol and identifies suitable machine learning techniques for the task. Chapter 3 leverages the ME-TIME dataset to demonstrate the potential of smartwatch-derived acceleration–deceleration curves as informative morphology-based HRV measure in machine learning models. We further explore a neural network architecture that incorporates per-subject, activity-based calibration using step count data, weekly aggregation of model outputs, and contrastive learning within a weakly supervised framework.

Chapters 4 and 5 address the critical challenge of inter-subject variability and propose strategies to enhance model generalizability across diverse populations. Both approaches build on the idea that knowing which heart rate measurement belongs to which subject can be leveraged to guide model adaptation. In Chapter 4, we group similar subjects and train separate machine learning models tailored to each group. In Chapter 5, we investigate factorization models that incorporate subject alignment through specialized contrastive and triplet loss functions, enabling the model to disentangle subject-specific patterns from condition-related signals.

Finally, chapter 6 offers a discussion of the findings, including insights from all chapters and highlighting their implications for future research and clinical deployment of wearable-based cardiovascular monitoring into real-world healthcare settings.

REFERENCES

[1] M. B. Vigliani. “Keeping a finger on the pulse—a brief history of cardiovascular medicine”. In: *Emerging Technologies for Heart Diseases*. Elsevier, 2020, pp. 1–22.

[2] M. A. Chamsi-Pasha and H. Chamsi-Pasha. “Avicenna’s contribution to cardiology”. In: *Avicenna Journal of Medicine* 4.01 (2014), pp. 9–12.

[3] M. M. Zarshenas, Z. Abolhassanzadeh, P. Faridi, and A. Mohagheghzadeh. “Sphygmology of ibn sina, a message for future”. In: *Heart Views* 14.3 (2013), pp. 155–158.

[4] O. C. Gruner. *A Treatise on the Canon of Medicine of Avicenna: Incorporating a Translation of the First Book*. United Kingdom: Classics of Medicine Library, 1984.

[5] I. Sina. *Rag Shenasi ya Resaleh dar Nabz (Pulsology, or Treatise on Pulse)*. Tehran, Iran: Selsele Intisharat-e Anjomane Asare Melli, 1951.

[6] D. Biswas, N. Simões-Capela, C. Van Hoof, and N. Van Helleputte. “Heart rate estimation from wrist-worn photoplethysmography: A review”. In: *IEEE Sensors Journal* 19.16 (2019), pp. 6560–6570.

[7] D. R. Bassett, L. P. Toth, S. R. LaMunion, and S. E. Crouter. “Step counting: a review of measurement considerations and health-related applications”. In: *Sports Medicine* 47 (2017), pp. 1303–1315.

[8] F. Shaffer, R. McCraty, and C. L. Zerr. “A healthy heart is not a metronome: an integrative review of the heart’s anatomy and heart rate variability”. In: *Frontiers in psychology* 5 (2014), p. 1040.

[9] F. Shaffer and J. P. Ginsberg. “An overview of heart rate variability metrics and norms”. In: *Frontiers in public health* 5 (2017), p. 258.

[10] P. K. Stein, M. S. Bosner, R. E. Kleiger, and B. M. Conger. “Heart rate variability: a measure of cardiac autonomic tone”. In: *American heart journal* 127.5 (1994), pp. 1376–1381.

[11] M. Malik and A. J. Camm. “Heart rate variability”. In: *Clinical cardiology* 13.8 (1990), pp. 570–576.

[12] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri. “Heart rate variability: a review”. In: *Medical and biological engineering and computing* 44 (2006), pp. 1031–1051.

[13] R. S. Wijesurendra and B. Casadei. “Mechanisms of atrial fibrillation”. In: *Heart* 105.24 (2019), pp. 1860–1867.

[14] S. H. Kim, K. R. Lim, J.-H. Seo, D. R. Ryu, B.-K. Lee, B.-R. Cho, and K. J. Chun. “Higher heart rate variability as a predictor of atrial fibrillation in patients with hypertension”. In: *Scientific Reports* 12.1 (2022), p. 3702.

[15] A. A. Khan, G. Y. Lip, and A. Shantsila. “Heart rate variability in atrial fibrillation: The balance between sympathetic and parasympathetic nervous system”. In: *European journal of clinical investigation* 49.11 (2019), e13174.

[16] C. D. Kemp and J. V. Conte. “The pathophysiology of heart failure”. In: *Cardiovascular pathology* 21.5 (2012), pp. 365–371.

[17] R. E. Kleiger, J. P. Miller, J. T. Bigger Jr, and A. J. Moss. “Decreased heart rate variability and its association with increased mortality after acute myocardial infarction”. In: *The American journal of cardiology* 59.4 (1987), pp. 256–262.

[18] S. Guzzetti, R. Magatelli, E. Borroni, and S. Mezzetti. “Heart rate variability in chronic heart failure”. In: *Autonomic neuroscience* 90.1-2 (2001), pp. 102–105.

[19] J. Park, H. S. Seok, S.-S. Kim, and H. Shin. “Photoplethysmogram analysis and applications: an integrative review”. In: *Frontiers in physiology* 12 (2022), p. 808451.

[20] M. Nitzan, A. Romem, and R. Koppel. “Pulse oximetry: fundamentals and technology update”. In: *Medical Devices: Evidence and Research* (2014), pp. 231–239.

[21] M. Nitzan, I. Faib, and H. Friedman. “Respiration-induced changes in tissue blood volume distal to occluded artery, measured by photoplethysmography”. In: *Journal of biomedical optics* 11.4 (2006), pp. 040506–040506.

[22] T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology. “Heart rate variability: standards of measurement, physiological interpretation, and clinical use”. In: *Circulation* 93.5 (1996), pp. 1043–1065.

[23] R. Sassi, S. Cerutti, F. Lombardi, M. Malik, H. V. Huikuri, C.-K. Peng, G. Schmidt, Y. Yamamoto, D. Reviewers: B. Gorenek, *et al.* “Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the European Heart Rhythm Association co-endorsed by the Asia

Pacific Heart Rhythm Society". In: *Ep Europace* 17.9 (2015), pp. 1341–1353.

[24] C. M. DeGiorgio, P. Miller, S. Meymandi, A. Chin, J. Epps, S. Gordon, J. Gornbein, and R. M. Harper. "RMSSD, a measure of vagus-mediated heart rate variability, is associated with risk factors for SUDEP: the SUDEP-7 Inventory". In: *Epilepsy & behavior* 19.1 (2010), pp. 78–81.

[25] D. J. Ewing, J. Neilson, and P. Travis. "New method for assessing cardiac parasympathetic activity using 24 hour electrocardiograms." In: *Heart* 52.4 (1984), pp. 396–402.

[26] P. Hämmerle, C. Eick, S. Blum, V. Schlageter, A. Bauer, K. D. Rizas, C. Eken, M. Coslovsky, S. Aeschbacher, P. Krisai, *et al.* "Heart rate variability triangular index as a predictor of cardiovascular mortality in patients with atrial fibrillation". In: *Journal of the American Heart Association* 9.15 (2020), e016075.

[27] A. H. Khandoker, C. Karmakar, M. Brennan, M. Palaniswami, and A. Voss. "Nonlinear Methods of HRV: The Poincaré Plot". In: *Poincaré Plot Methods for Heart Rate Variability Analysis*. Springer, 2013. Chap. 2, pp. 13–23.

[28] T. Ziemssen and T. Siepmann. "The investigation of the cardiovascular and sudomotor autonomic nervous system—a review". In: *Frontiers in neurology* 10 (2019), p. 53.

[29] K. Li, H. Rüdiger, and T. Ziemssen. "Spectral analysis of heart rate variability: time window matters". In: *Frontiers in neurology* 10 (2019), p. 545.

[30] C. W. Myers, M. A. Cohen, D. L. Eckberg, and J. A. Taylor. "A model for the genesis of arterial pressure Mayer waves from heart rate and sympathetic activity". In: *Autonomic Neuroscience* 91.1-2 (2001), pp. 62–75.

[31] S. Elsenbruch, Z. Wang, W. C. Orr, and J. Chen. "Time-frequency analysis of heart rate variability usingshort-time Fourier analysis". In: *Physiological Measurement* 21.2 (2000), p. 229.

[32] D. Chemla, J. Young, F. Badilini, P. Maison-Blanche, H. Affres, Y. Lecarpentier, and P. Chanson. "Comparison of fast Fourier transform and autoregressive spectral analysis for the study of heart rate variability in diabetic patients". In: *International journal of cardiology* 104.3 (2005), pp. 307–313.

[33] M. Estévez, C. Machado, G. Leisman, T. Estévez-Hernández, A. Arias-Morales, A. Machado, and J. Montes-Brown. "Spectral analysis of heart rate variability". In: *International Journal on Disability and Human Development* 15.1 (2016), pp. 5–17.

[34] V. Pichot, J.-M. Gaspoz, S. Molliex, A. Antoniadis, T. Busso, F. Roche, F. Costes, L. Quintin, J.-R. Lacour, and J.-C. Barthélémy. “Wavelet transform to quantify heart rate variability and to assess its instantaneous changes”. In: *Journal of Applied Physiology* 86.3 (1999), pp. 1081–1091.

[35] S. M. Pincus. “Approximate entropy as a measure of system complexity.” In: *Proceedings of the national academy of sciences* 88.6 (1991), pp. 2297–2301.

[36] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger. “Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series”. In: *Chaos: an interdisciplinary journal of nonlinear science* 5.1 (1995), pp. 82–87.

[37] G. M. Felker, P. S. Cuculich, and M. Gheorghiade. “The Valsalva maneuver: a bedside “biomarker” for heart failure”. In: *The American journal of medicine* 119.2 (2006), pp. 117–122.

[38] J. Alvarez-Ramirez, J. Echeverria, M. Meraz, and E. Rodriguez. “Asymmetric acceleration/deceleration dynamics in heart rate variability”. In: *Physica A: Statistical Mechanics and its Applications* 479 (2017), pp. 213–224.

2

MACHINE LEARNING FOR CARDIOVASCULAR OUTCOMES FROM WEARABLE DATA: SYSTEMATIC REVIEW FROM A TECHNOLOGY READINESS LEVEL POINT OF VIEW

Wearable technology has the potential to improve cardiovascular health monitoring using machine learning. It enables remote health monitoring and allows for diagnosis and prevention. In addition to detection of cardiovascular disease, it can exclude a diagnosis in symptomatic patients, preventing unnecessary hospital visits. Furthermore, early warning systems can aid the cardiologist in timely treatment and prevention.

This study aims to systematically assess the literature on detecting and predicting outcomes of patients with cardiovascular diseases by using machine learning with data obtained from wearables to gain insights into the current state, challenges, and limitations of this technology.

We searched PubMed, Scopus, and IEEE Xplore on September 26, 2020, with no restrictions on the publication date and by using keywords such as "wearables", "machine learning," and "cardiovascular disease." Methodologies were categorized and analyzed according to machine

This chapter has been published in JMIR Medical Informatics **10**, 1 (2022).

learning–based technology readiness levels (TRLs), which score studies on their potential to be deployed in an operational setting from 1 to 9 (most ready).

We searched PubMed, Scopus and IEEE Xplore on September 26, 2020 with no restrictions on the publication date using keywords: wearables, machine learning and cardiovascular disease. Methodologies were categorized and analyzed according to machine learning based technology readiness levels (TRLs) that score studies on their potential to be deployed in an operational setting from 1 to 9 (most ready).

After the removal of duplicates, application of exclusion criteria, and full-text screening, 55 eligible studies were included in the analysis, covering a variety of cardiovascular diseases. We assessed the quality of the included studies and found that none of the studies were integrated into a health care system (TRL<6), prospective phase 2 and phase 3 trials were absent (TRL<7 and 8), and group cross-validation was rarely used. These issues limited these studies' ability to demonstrate the effectiveness of their methodologies. Furthermore, there seemed to be no agreement on the sample size needed to train these studies' models, the size of the observation window used to make predictions, how long participants should be observed, and the type of machine learning model that is suitable for predicting cardiovascular outcomes.

Although current studies show the potential of wearables to monitor cardiovascular events, their deployment as a diagnostic or prognostic cardiovascular clinical tool is hampered by the lack of a realistic data set and proper systematic and prospective evaluation.

2.1. INTRODUCTION

2.1.1. BACKGROUND

The use of diagnostic modalities in cardiovascular disease is often limited to hospital visits. As a result, the clinical value may be limited by the short observation period. This is especially problematic for cardiovascular problems that do not manifest constantly, such as paroxysmal arrhythmias, heart failure, or even chest discomfort that may not be present during the hospital visit. Advancements in eHealth, especially in wearable technology, such as electrocardiograms (ECGs)[1] and photoplethysmograms (PPGs)[2], and subsequent signal processing by machine learning have enabled new opportunities for remote monitoring in the outpatient setting.

Continuous monitoring over long periods has shown to be effective[3, 4]. For example, remote monitoring of patients with cardiac diseases, using pacemakers or implantable cardioverter defibrillators and patients with heart failure have improved patient care[5]. However, current sensors used in health care, such as Holter devices, are limited to a maximum of 14 days (but typically endure 24 hours) of continuous monitoring, limiting the use of these devices. Overcoming this could enable early warning systems for acute events such as cardiac arrest and could capture subtle cardiovascular exacerbation or rehabilitation that manifests over a much longer time because of, for example, changes in lifestyle or intervention.

Although widely used, currently 24-hour ECG or blood pressure monitoring devices are cumbersome to wear and impose a burden on patients in a longitudinal setting. Rechargeable, easy-to-wear sensors, such as smartwatches, are becoming an interesting alternative as they contain sensors with a potentially unlimited observation period with minimal burden to the patient for a fraction of the costs. However, the signals that these wearables measure, such as the PPG-derived heart rate, activity, and skin temperature, are not clinically informative enough for clinical decision-making by a cardiologist. With current developments in artificial intelligence (AI), a powerful solution is expected from machine learning algorithms that can learn the relationship between the wearable sensor signals and a cardiovascular outcome in a (fully) data-driven manner.

Another great benefit of automatic cardiovascular diagnostics and prognostics by machine learning is minimizing inter- and intraobserver variability, which is a major problem in the subjective interpretation of clinical and diagnostic information by human cardiologists. Interobserver disagreement[6, 7] because of, for example, differences in experience or specialization and intraobserver disagreement because of stress or fatigue[8], can be minimized. Variations in clinical practice may lead to medical errors, whereas automatic systems are not (or less) susceptible to such factors. Another possibility is to exclude patients who experience

symptoms such as chest pain, which are not caused by cardiovascular disease. Automatic exclusion of these patients can reduce unnecessary visits to a cardiologist; relieving the cardiologist, thereby increasing the capacity of cardiovascular care; and directing patients to the proper specialist quicker.

Because of these promises, the field of research on diagnosing cardiovascular events from wearable data is very active and many machine learning solutions are being presented to automatically detect cardiovascular events. Various reviews have been presented to categorize the developed machine learning tools. A study by Krittanawong et al.[9] shows that a plethora of wearable devices are researched for a variety of cardiovascular outcomes and discusses a paradigm for remote cardiovascular monitoring consisting of sensors, machine learning diagnosis, data infrastructure, and ethics. They conclude that especially the latter two aspects have several unaddressed challenges. An overview of wearable devices on the market is provided by Bayoumy et al.[10]. The study reports their frequency of use in (cardiovascular) trials and Food and Drug Administration status. As reported by Giebel and Gissel[11], most mobile health devices for atrial fibrillation detection are not Food and Drug Administration approved and therefore cannot be used in cardiovascular monitoring systems.

2.1.2. OBJECTIVES

Although many machine learning tools have been proposed and studies have shown good performance, they do not seem to have been implemented in operational and functional health care systems. Therefore, we decided to systematically review the machine learning tools to detect cardiovascular events from wearable data from the perspective of their technology readiness level (TRL), that is, how far these proposed tools are in realizing an operational system and what factor is impeding them to get there. The TRL paradigm originates from the National Aeronautics and Space Administration and is a way to assess the maturity level of a particular technology used in space travel by giving solutions a score from 1 to 9 in increasing order of readiness, from basic technology research (score 1) to launch operations (score 9)[12].

Interestingly, 2 studies tailor the TRL framework for medical machine learning. A study by Komorowski[13] proposes a TRL for supervised, unsupervised, and reinforcement learning problems and describes criteria to reach TRLs 3, 4, 6, and 7. A description of the 9 TRLs for medical machine learning in intensive care medicine, including examples, is proposed by Fleuren et al.[14]. We review the wearable-based cardiovascular machine learning solutions following the framework by Fleuren et al.[14] adjusted for remote medicine. We

identify aspects in the studies and systematically assign these to TRLs and group some of the TRLs together in a taxonomy to help interpret their relevance (Figure 2.1). We address the overuse of benchmark data sets, considerations on data acquisition related to the environment and type of sensor, integration in a health care system, construction of the machine learning model, and subsequent model validations.

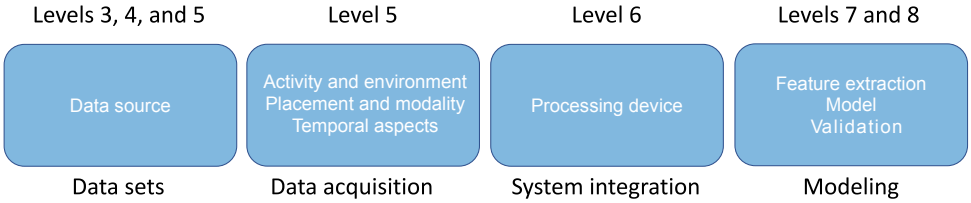


Figure 2.1.: Taxonomy of the eligible studies. TRLs are based on the proposed descriptions for machine learning for medical devices proposed by Fleuren et al[14]. The studies were categorized according to the relevance of their content to these descriptions (aspects within boxes) and were grouped and assigned to the different TRLs (below and above boxes). TRL: technology readiness level.

By assessing current methods by their technological readiness, we show that the current methodologies are promising but that deployment is severely hampered by the lack of realistic data sets and proper systematic and prospective evaluation. To arrive at a readiness that is operational at the health care system level, these bottlenecks need to be resolved.

2.2. METHODS

2.2.1. SCREENING

The systematic review was performed by following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines[15], as shown in Figure 2.2. We followed the patient or population, intervention, comparison, and outcomes framework for our research question, which was as follows: “In patients with cardiovascular disease, using machine learning with data from wearables, what methods and accompanying limitations are used, to deploy this technology to detect and predict cardiovascular disease in standard healthcare?”

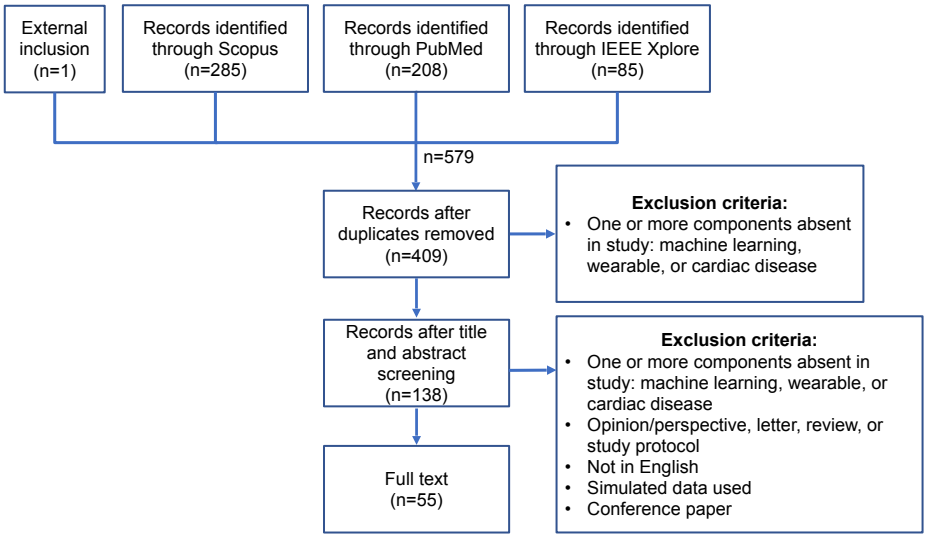


Figure 2.2.: PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the systematic review.

2.2.2. STUDY INCLUSION

Search queries were performed on September 26, 2020, in the electronic databases Scopus, PubMed, and IEEE Xplore. Only peer-reviewed journals were considered. Studies were eligible for inclusion if data were acquired from wearables, a machine learning method was used, and had the goal to detect or predict cardiovascular disease (see Appendix A.1 for used queries). The following exclusion criteria were used: opinion or perspective, letter, review, study protocol, or conference paper; studies not in English; and studies in which only simulated data were used. The eligibility assessment was performed by the first author, ANJ. First, the title and abstract of each study were assessed for relevance based on the inclusion and exclusion criteria. The full texts of the remaining studies were then read and again subjected to the selection criteria. The second author, DT, verified this by reading a subsample of the selection.

2.2.3. TRL AND TAXONOMY

From the eligible studies through discussions with all authors, the first author, ANJ, identified some general overarching evaluation aspects that the studies had in common and assigned these studies to a taxonomy (Appendix A.2 [16-70]). These aspects were related to one or more TRLs, as defined by Fleuren et al [14]. Accordingly, the eligible studies were assigned to the taxonomy and different TRLs (2.1). The TRL

framework states that studies that use only a benchmark data set as a data source do not progress further than level 3. Furthermore, the framework originally grouped levels 3 and 4 together. We split these levels and assigned studies using their own acquired data without an external validation set from a different study level 4.
Next, we assigned studies that use an external validation set from a different study to level 5; although, according to Fleuren et al [14], level 5 further requires that the acquired data set is realistic. However, we interpreted the independently acquired data representative of data recorded during the deployment of the machine learning system as realistic. Therefore, we differentiated levels 3, 4, and 5 mostly on the data sets being used for model deployment and related these levels to the data sets taxonomy. As level 5 mainly focuses on realistic data sets we also assigned practical aspects of the wearables to this TRL.
Here, we differentiated the following three aspects: (1) which modality is being measured by the wearable and where on the body it is placed; (2) under which conditions data are measured, such as in the wild or in controlled environments; and (3) for how long data are recorded, that is, the temporal aspect of the acquired data. Level 6 required integrating the machine learning model into a health care system. Therefore, the device in which the model was integrated into was assigned to this level. Finally, levels 7 and 8 required demonstrating the model as a cardiovascular tool. Therefore, the model effectiveness and validation aspects were assigned to these levels. Levels 1, 2, and 9 were disregarded here because none of the papers fit into these categories.

2.3. RESULTS

2.3.1. ARTICLE IDENTIFICATION

A total of 578 records were retrieved from electronic databases. After the removal of duplicates, 70.8% (409/578) of records remained. One was externally included as it fulfilled the inclusion criteria but was missed by the search query because it did not explicitly mention the term machine learning. As shown in Figure 2.2, these were further narrowed down during title or abstract screening, resulting in 23.9% (138/578) of records. Finally, after full-text reading, 9.5% (55/578) of records remained to be covered in this study.
We related each of the studies to different TRLs for machine learning methods (*Methods*) according to an identified taxonomy of different evaluation criteria that relate to these TRLs (Figure 2.1; *Methods*). The TRL framework states that studies that use only a benchmark data set do not progress further than level 3.

2.3.2. STUDY CHARACTERISTICS

The key characteristics of the eligible studies are summarized in Appendix A.2. Notably, of the 55 studies, 27 (49%) exclusively used benchmark data sets, which were all defined as benchmark studies. Furthermore, of the 55 included studies, 6 (11%) were published before 2018 and the remaining 49 (89%) were published thereafter. In the following sections, the study characteristics are described more closely based on the taxonomy.

2.3.3. ACTIVITY AND ENVIRONMENT (LEVEL5)

For studies that did not use benchmark data sets, they reported the data acquired either in a controlled environment (hospital or research laboratory) or in a free-living environment, where participants were remotely observed performing their natural daily routines. The latter is also known as in-the-wild. Furthermore, the activities of the participants can be divided into sedentary or active during data acquisition. To capture these two related aspects, we assigned studies on an axis representing a controlled environment and sedentary activity on one side and in-the-wild measurement of active participants on the other side of the axis (Figure 2.3). Interestingly, only 5 [16][17][18][19][20] studies mapped to the active, free-living situation that complied with the requirement of realistic data acquisition for these aspects that map to TRL 5. Thus, only one-tenth of the studies used the potential of wearables to be used for remote, longitudinal monitoring.

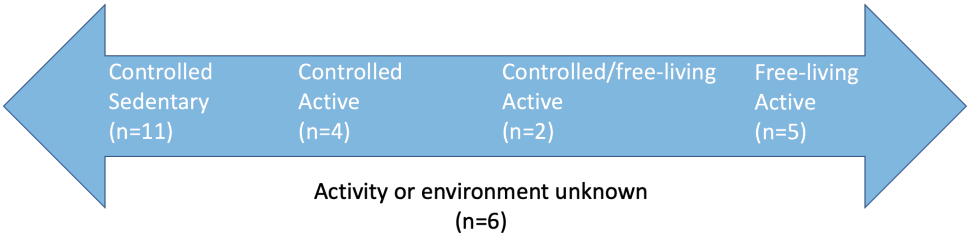


Figure 2.3.: Studies ordered based on participant activity and acquisition environment. The leftmost scenario indicates highly controlled acquisition with sedentary participants. The opposite is described by the rightmost scenario where participants are monitored in an active, free-living situation. Controlled environment includes hospitals or laboratories. Free-living participants are monitored during their daily routines.

2.3.4. PLACEMENT AND MODALITY (LEVEL 5)

Realistic data acquisition requires continuous monitoring. Practically, the wearable should therefore not burden the participant when wearing. This burden depended mostly on the placement of the sensor on the body. In addition, the placement also restricted the type of biometric signals that could be measured, which was referred to as the modality. We categorized studies based on the placement and modality for the nonbenchmark studies jointly (Figure 2.4). The sensor placements for cardiovascular monitoring that results in the least burden for the patient, and thus would be the best candidates to acquire a realistic data set, were the wrist and finger. Less than half (N=13) of the studies were reported with such placements, of which 8 (62%) studies acquired one modality: 3 (23%) studies acquired wrist-based ECGs [18][21][22], 2 (15%) studies acquired wrist-based PPGs [17][23], and 3 (23%) studies acquired finger-based PPGs [24][25][26]. Of the 13 studies, the remaining 5 (39%) studies acquired wrist-based multimodal data: 4 (31%) studies acquired PPGs and accelerometer data [19][20][27][28] and 1 (8%) study acquired both ECGs and PPGs [29]. Thus, the wrist and finger severely limited the additional modalities that were measured (usually only acceleration), although wearables were shown to be able to measure increasing number of modalities [10].

2.3.5. TEMPORAL ASPECTS (LEVELS 5,7, AND 9)

Besides the requirement of a realistic data set in level 5, levels 7 and 8 required phase 2 and phase 3 studies, respectively. In the context of drug testing, this requires an investigation of the effective, but safe, drug dosage. Analogously, for wearable machine learning, this translated to the time a participant must be exposed to a machine learning model before a cardiovascular outcome could be accurately detected or predicted. Therefore, a realistic deployment setting is dependent on the length of time participants are observed.

As it is further essential to characterize the data for reproducibility and the description under which circumstances a model is valid, we decided to outline the temporal aspect of the acquired wearable data in more detail. We recognized the following four levels of time aspects: (1) study duration, (2) observation period, (3) recording duration, and (4) input window size (Figure 2.4). Within the study duration, patients were included and observed for a certain period—the observation period. The lengths of these periods had an impact on the realistic deployment of a system. For example, Quer et al [30] used wrist-worn Fitbit devices to show that resting heart rate within individuals had a significant seasonal trend in longitudinal data. Therefore, a model constructed using data from a certain period might not be valid for another period. It was therefore important to consider how long the participants were observed

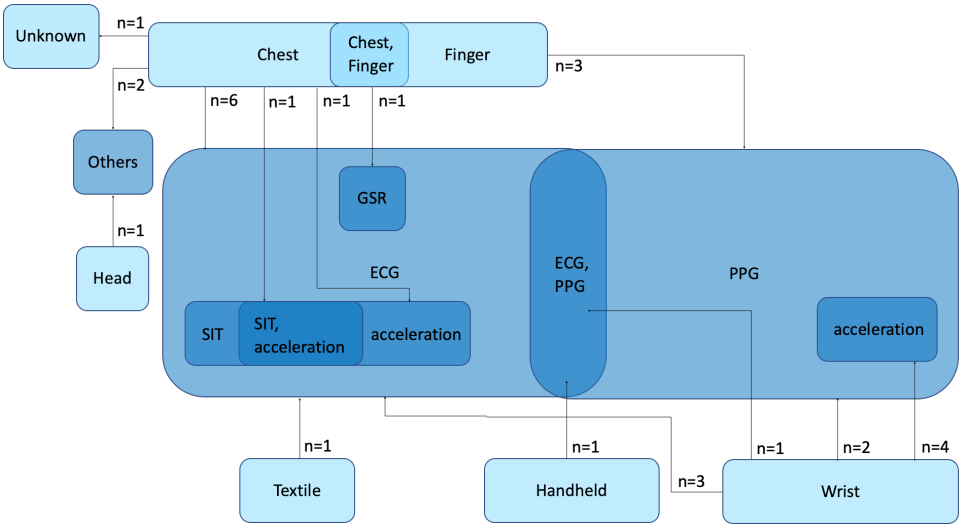


Figure 2.4.: Placement and modalities of wearable sensors: light blue, placement of sensors; blue, modalities used. Others: head, near-infrared spectroscopy; chest, seismocardiography or gyrocardiography. Overlapping blocks represent multiple placements or modalities used. ECG: electrocardiogram; GSR: galvanic skin response; PPG: photoplethysmogram; SIT: skin impedance and temperature.

to ensure this seasonal effect was incorporated in the model. Within the observation period, the wearable recorded a time series. Theoretically, this could be as long as the observation period itself. However, patients could interrupt the measurements for several reasons (e.g., to charge the device and low compliance rate). We denoted the duration of a continuously measured part of the time series as the recording duration. Finally, the records were further segmented into windows, from which features were generated or which were used as raw inputs to a machine learning model. We referred to the duration of these windows as the input window size (I). We assessed the temporal aspects of all the non-benchmark studies (Figure 2.5). One study did not report any aspects 2.5 and was omitted from Figure 2.5. Another study used multiple fixed input window sizes to incorporate different timescales of the data [19]. Overall, most studies did not report all the aspects and were thus not comprehensive about their data characteristics. In almost all studies, the recording rate and input window size were reported, whereas the study and observation periods were mentioned in about half of the studies. For a realistic data set, required for level 5 and progression to level 7 or 8, the observation period and recording

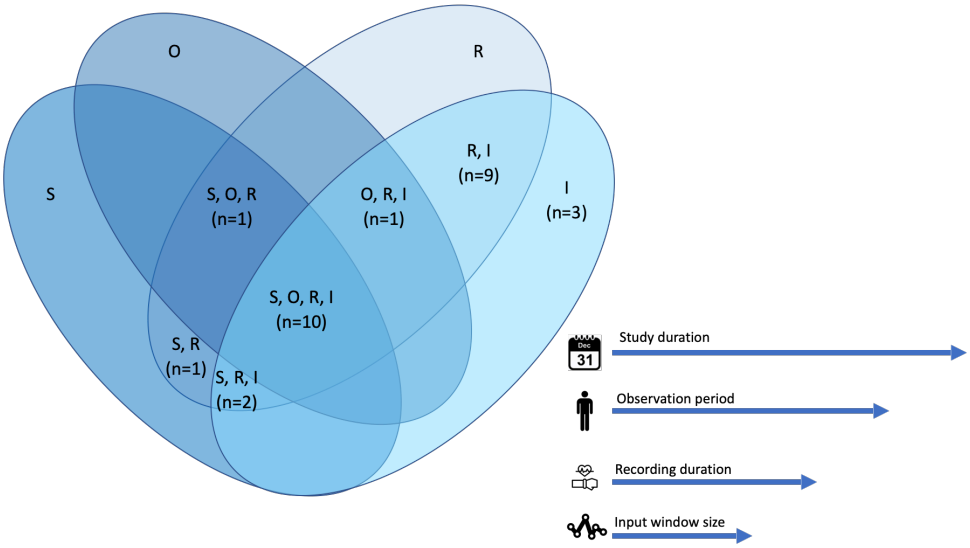


Figure 2.5.: Venn diagram of reported temporal aspects described in the studies. The S, O, R, and I are represented in the legend. I: input window size; O: observation period; R: recording duration; S: study duration.

duration were specifically important, as we found in 12 studies. Three studies used an observation period of 24 hours [23][31][32]; one for a week [17], one for 2 weeks [33], and one for 90 days [16]. Overall, 2 studies implied an observation period of months but did not explicitly report it [19][20]. One considered recordings of at least eight hours [19] and one reported an average recording duration of 11.3 hours [20]. Finally, only one [33] fully used the potential of wearables and reported a (near-) continuous recording duration.

2.3.6. CARDIOVASCULAR OUTCOMES (ALL LEVELS)

Although the required observation period and recording duration to detect or predict a cardiovascular outcome is still an open and active research topic, these periods will be different for different outcomes. Therefore, we inventoried which (combinations of) cardiovascular outcomes were considered in which studies (Figure 2.6). Interestingly, the control group was defined differently in each study. Only half of the non-benchmark studies included a (normal) sinus rhythm class as control and could therefore exclude the presence of cardiovascular disease in participants. From these, 8 studies [17][21][22][23][34][27][25][35] used data from healthy individuals to represent normal sinus rhythm. The remaining 6 studies [31][36][37][38][39][26] derived normal sinus

rhythm data from patients with arrhythmia (such as paroxysmal atrial fibrillation) or were unclear about the control group. Three studies had cardiovascular (disease) prevention as the target. One of these described this as a cardiovascular risk assessment where the predicted classes were healthy, precaution, and critical status [34]. Another study predicted vascular age and 10-year cardiovascular disease risk [37]. The third assigned a cardiorespiratory fitness score [33]. Notably, only the first 2 studies constructed a prognostic model. Two other prognostic models forecast cardiac arrest and heart failure exacerbation by forecasting rehospitalization after heart failure admission [16][21].

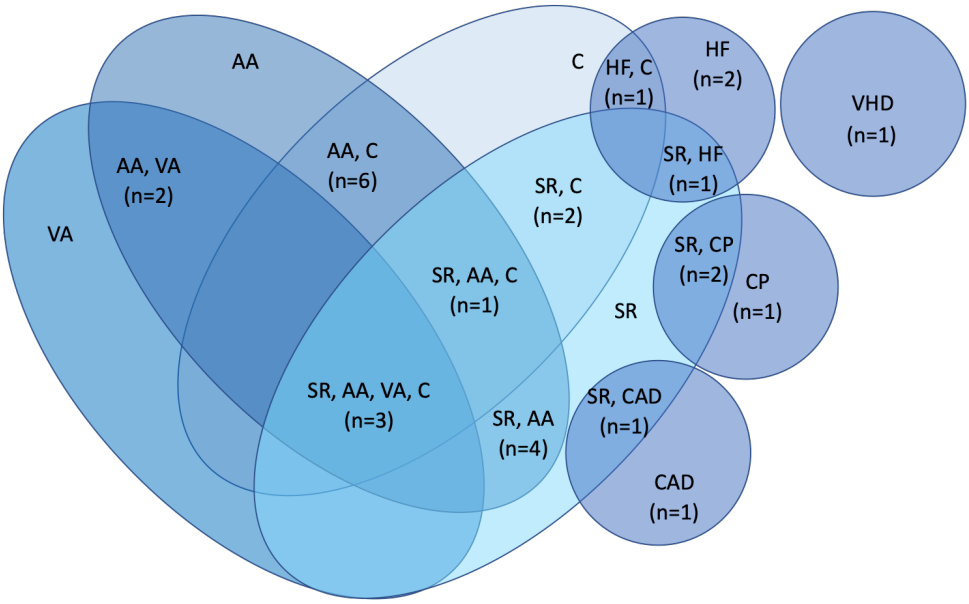


Figure 2.6.: Studies categorized according to the type of cardiovascular outcomes predicted by the models. AA: atrial arrhythmia; C: control; CAD: coronary artery disease; CP: cardiovascular prevention; HF: heart failure; SR: sinus rhythm; VA: ventricular arrhythmia; VHD: valvular heart disease.

2.3.7. BOTTLENECK TRL 5

Although many cardiovascular outcomes were investigated with wearables, the promising studies that have reached level 5 were all focused on atrial arrhythmia using wrist-based PPGs. However, their temporal properties were often inconclusive, as they were not reported. Moreover, to progress to level 6, a model should be functional within a health care system (even if it was merely used observationally). None of the studies progressed to this level. An overview of the level 5 models, including the modalities that they are based on, is given in Table 2.1. Although none of the methodologies progressed to level 6, we decided to prospectively evaluate the studies to investigate the progression of the current state.

Table 2.1.: Studies fulfilling requirements for technology readiness level 5. Shown are observation period (O), recording duration (R), input window size (I), photoplethysmogram (PPG), electrocardiogram (ECG) and not reported (NR).[1]:Sensor-provided heart rate and step counter data.

Study	Outcomes	Modality	O	R	I
Torres-Soto and Ashley [17]	Sinus rhythm, atrial arrhythmia	PPG	1 week	NR	25 seconds
Bashar et al [18]	Atrial arrhythmia, ventricular arrhythmia	ECG	NR	NR	2 minutes
Tison et al [19]	Atrial arrhythmia, control	PPG, accelerometer[1]	NR	>8 hours a day	5 seconds, 30 seconds, 5 minutes, and 30 minutes
Wasserlauf et al [20]	Atrial arrhythmia, control	PPG, accelerometer	NR	11.3 hours a day	1 hour

2.3.8. PROCESSING DEVICE (LEVEL 6)

Integration in a health care system could be carried out on different devices. These studies demonstrated their models on either a computer (eg, a server), smartphone, or embedded device (Table 2.2). Only the latter two enabled real-time cardiovascular monitoring locally on the patient side, required for real-time detection and prevention of acute cardiovascular disease, as real-time information exchange to an external system would require high battery consumption and was therefore not feasible. Smartphones were used in both benchmark [40][41][42] and non-benchmark [21][25][35][38] studies. Embedded devices, however, had only been demonstrated in benchmark studies [43][44][45][46].

Table 2.2.: Processing device of trained models used in studies.

Processing Device	Benchmarks Included, n	Benchmarks Excluded, n
Computer	44	24
Smartphone	7	4
Embedded device	4	0

2.3.9. FEATURE EXTRACTION METHODS (LEVELS 7 AND 8)

Levels 7 and 8 of the TRL assessed the model effectiveness through phases 2 and 3 clinical trials. We translated that to what features from the observed modalities were being used to construct the models. A significant number of studies used ECG as a modality and used different information from fiducial points [47] to extract features (Figure 2.7). In many studies, samples were selected before and after the R-peak. For example, the RR interval is the time interval between 2 adjacent R-peaks. Some studies also used techniques to locate other fiducial points and used the time interval between them as features [48]. Together, we denoted these types of features as waveform information features.

Next to the specific ECG features, more general features could be derived, such as statistical features (e.g., heart rate variability derived from 10 RR intervals) or spectral features obtained through techniques such as the Fourier transform. Raw data could also be used as features upon which a neural network can be used to automatically learn informative features [49]. Next to the features based on the sensed signal, demographic information could be used to provide more context [34][28]. Benchmark studies mostly use raw features (using the same data set) and were, therefore, excluded from this study. However, it is noteworthy that 2 of these used more advanced methods, namely, compressed learning [50] combined with dynamic time warping [51].

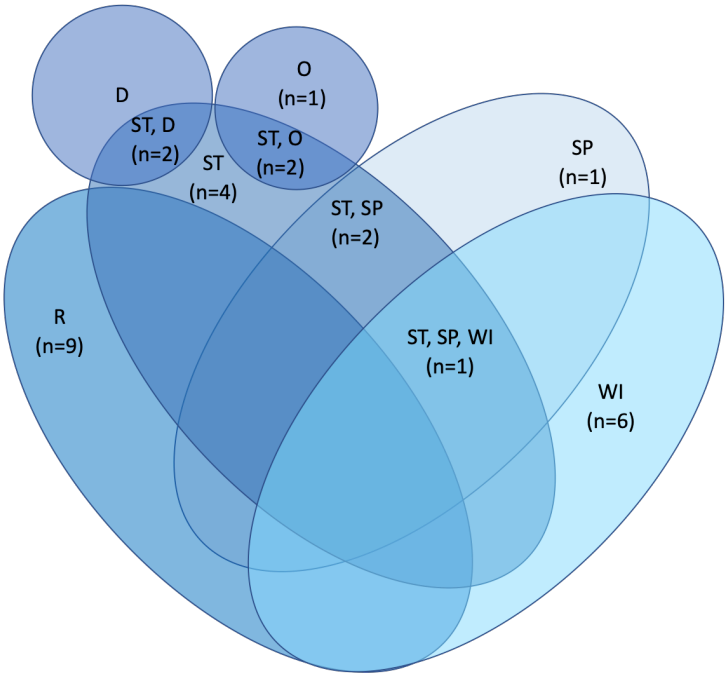


Figure 2.7.: Features used in the studies. D: demographic; O: others; R: raw; SP: spectral; ST: statistical; WI: waveform information.

The most commonly used features were raw features (studies: 9/28, 32.1%). This was followed by waveform information and statistical features. In all, 2 studies also included demographic metadata from participants [34][28]. One study used hemoglobin parameters [52], which we represented in the others group in Figure 2.7. Interestingly, 1 study included timestamps [19]. From the 11 studies that used multimodal data (Figure 2.4), 6 (55%) studies extract features for each modality were separately extracted. Of the 11 studies, the remaining 5 (46%) studies exploited the covariance among the modalities in feature extraction, although 1 (9%) study did not elaborate on the exact method [16]. For example, of the 15 studies, 1 (9%) study computed the time between an R-peak in the ECG and the closest following peak in the PPG [37]. Of the 5 studies, 2 (40%) studies concatenated windows of the different modalities and then extracted the features [20][53] and 1 (20%) study concatenated windows whereafter a convolutional layer in a neural network is used to automatically extract features from the concatenated data [19].

2.3.10. Model construction methods (levels 7 and 8)

Another aspect that defines the model effectiveness relates to the type of models being constructed, which we categorized across both the benchmark and nonbenchmark studies (Table 2.3). Most of the studies used a neural network, and most of them were nonsequential (e.g., convolutional and multilayer perceptron). A noteworthy type is the spiking neural network [54][55], which is designed to be energy efficient and suitable for real-time cardiovascular monitoring in an embedded device. Although sequential models were specifically designed for sequence or time series, these types of models were used much less. Some studies had combined sequential and nonsequential neural network architectures [17][19][31][44][49][56]. After the neural networks, most of the models were classical machine learning methods, including linear models: support vector machines; decision trees; and similarity-based models, such as k-nearest neighbor classifiers. Furthermore, ensemble methods had been used that combined multiple simpler models to construct a more complex model [22][34][46][53][57][58][59]. Finally, 2 studies used models that explicitly exploit the hierarchical structure of medical time series data: a hierarchical Bayesian model [33] and a Multiple-Instance Learning via Embedded instance Selection model [23].

Table 2.3.: Types of machine learning models used in the studies.

Model Type	Number of Times Used
Nonsequential	30
Classical	20
Ensemble	9
Sequential neural network	6
Nonsequential + sequential neural network	5
Hierarchical	2

2.3.11. Validation (levels 7 and 8)

The effectiveness of a model was heavily influenced by the number of samples with which the model had been trained. In phase 2 and phase 3 studies, a priori power analyses were performed to estimate the required sample size per group or class to observe an effect. It was empirically shown by Quintana [60] that for heart rate variability studies, an effect size of 0.25, 0.5, and 0.9 corresponded to a low, medium, and high effect, respectively. The corresponding sample sizes were 233, 61, and 21 for 80% statistical power and 312, 82, and 28 for a 90% statistical power. We considered nonbenchmark studies

with a sufficient sample size per group or class, from which 9 studies remained. From the remaining 9 studies, a power of 90% was reached with small [19][20][24] and large [16][25][26][28] effect sizes, and 2 studies [27][31] achieved 80% power with a large effect size. This showed that studies generally choose a train sample size (per group or class) that is too small to find a significant effect based on a priori power analysis.

In contrast to a priori power analysis, the purpose of model validation is to retrospectively analyze the performance of the model on data it has not seen before, that is, to assess the generalization error of the model. The included studies chose from 2 validation schemes: cross-validation and holdout [61] (Figure 2.8), although 5 studies [16][20][34][32][62] did not report the validation method. When splitting data into training and testing, one needed to ensure non-overlapping grouping and stratification of the data (Figure 2.8). With non-overlapping grouping [63], one ensured that the same group of data did not appear in both the training and test sets, for example, avoiding that data from the same participant was in both the training and test set, albeit the samples might be from different periods. With stratification, one ensured that both the training samples and the test samples exhibit a similar proportion of samples for an arbitrary variable. For example, it was important to keep the proportion of men and women consistent or to ensure that the proportion of sensor samples representing normal rhythm and arrhythmia is equal. For progressing to TRL 7, 4 studies used leave-one-subject-out group cross-validation [18][23][33][48] and 4 other types of group cross-validation [27][25][26][46]. Ideally, a stratified group cross-validation is used, but none of the studies used this. In addition to validation strategies, it is important to use replication data, that is, completely independently acquired data, which was only done in 11 [17][18][21][24][29][35][36][38][39][42][64] studies.

It is important to realize that data sets could suffer from highly imbalanced classes. An example is when there are proportionally more samples representing sinus rhythm than atrial fibrillation. In this case, the model may be biased to focus more on correctly classifying sinus rhythm, as this contributed more to higher overall classification performance. However, this led to poor characterization of cardiovascular disease, as the corresponding samples would be misclassified more often than sinus rhythm. In all, 6 studies [31][43][65][66][67][68] mitigated this by (randomly) up-sampling the minority class. A total of 4 studies [22][27][50][55] used the synthetic minority oversampling technique [69].

Finally, it is noteworthy that some studies [43][44][45][48][51][54][70] constructed a semi–patient-specific model. This could be beneficial, as there were large differences in heart rate data among individuals [30]. This was done by training only a small number of samples from

the target patient together with data from other patients. The test set consisted of the remainder of the target patient's samples, which caused overlapping grouping between the training and test sets.

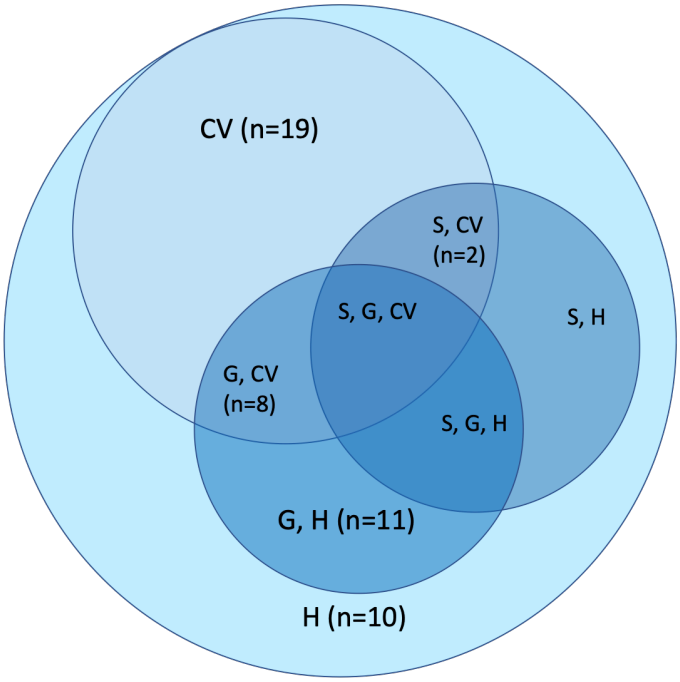


Figure 2.8.: Venn diagram of validation methods used in the studies. CV: cross-validation; G: grouping; H: holdout; S: stratification.

2.4. DISCUSSION

2.4.1. PRINCIPAL FINDINGS

We have shown that machine learning–based technologies that detect cardiovascular outcomes using wearables, bottleneck at TRL5, most dominantly on the requirement of proper realistic data acquisition. To progress to the next level of technology readiness, models need to become operational (either interventional or observational) in a health care system. A study by Komorowski [13] supports these observations and defines the lack of testing or deployment in clinical practice as an information bottleneck, which often occurs in medical machine learning. Moreover, half of the eligible studies used a benchmark data set (27/55, 49%), and the most common data set [71] was used 18 times. We argue that overusing a data set can introduce bias and overfitting, effectively making such a data set useless, thereby increasing the need for realistic

data sets even more.

The usefulness of wearable cardiovascular diagnostics lies in free-living and active situations because the low burden for wearing them and the 24/7 monitoring abilities. Placement of the sensor on the wrist does fit these criteria best. Moreover, commercial-grade smartwatches can measure multimodal data with low battery consumption. This makes these types of sensors promising to use wearable technology for cardiovascular diagnostics. However, most studies do not fully demonstrate this potential. Moreover, very few prognostic models have been proposed so that cardiovascular disease prevention using wearable machine learning is, in fact, not (yet) well researched.

Although most studies include detailed baseline characteristics of the study population, it is worrisome that the data were not described with a similar level of consistency, structure, and detail. For example, some studies (explicitly or implicitly) have reported acquiring continuous wearable data, but participants do need to take off the device for charging or otherwise have a low compliance rate. These studies then fail to report these details; thus, it is unknown how continuous the data, that is, the length of the recording duration, actually is. We believe that, analogous to the baseline characteristics, data characteristics should be reported in detail to predict how a model will generalize when deployed in a particular setting and environment.

The segmentation of the time series data in the windows was performed with a fixed window size in all studies. None of the studies have considered a variable-length or adaptive window size. Furthermore, no previous physiological knowledge has been used to determine informative timescales. For example, the exercise-recovery curve (usually obtained from an exercise tolerance test) is often used to quantify cardiovascular characteristics during activity. This describes a participant's ability to adaptively increase the heart rate during exercise and recover it back to a resting level after exercise. Studies that had access to accelerometer data did not look at similar timescale events. To this end, we believe that identifying informative timescales within the time series and incorporating this in the model can be valuable to detect cardiovascular diseases.

Remarkably, studies primarily prefer nonsequential neural networks over sequential ones, although the latter is designed for time series data. Similarly, the hierarchical structure of the data has rarely been exploited in the published models. We advocate that much more emphasis should be on the exploration of these models, although this also requires larger data sets as these methods are data hungry.

Although some studies make use of a healthy control group, most do not include a group with no arrhythmia, sinus rhythm, or a similar group, although diagnosing a participant having no arrhythmia at all is just as, or even more powerful, than detecting a specific heart

problem. From a machine learning point of view, this can be seen as a one-class classification (outlier detection) problem: instead of predicting a diverse set of clinical outcomes, the focus of these models lies in modeling the normal class as good as possible and consider deviating data as abnormal. Thus, this would be an interesting avenue to explore. In general, it is important to have clearly defined data annotations. For example, some studies have annotated sinus rhythm events in patients with arrhythmia. One might question whether this is similar to annotated sinus rhythm events for nonarrhythmic individuals and whether a machine learning–based approach might fail by mixing these annotations.

We have shown that studies use a training sample size that is too small according to a priori power analysis. Sample size determination in machine learning [72] is focused on posthoc methods, such as learning curves [73]. Prehoc methods, such as power analysis, are difficult in machine learning, as there are many factors that influence the effect size of the model. Furthermore, we have discussed different validation schemes that can be used. An important observation is that a significant number of studies do not validate their model using a non-overlapping grouping strategy. We believe that validation based on nonoverlapping grouping is crucial for cardiovascular machine learning and any medical machine learning validation in general. Without, experiments will simply suggest performances that are too optimistic.

We have shown that only a few papers used multimodal data and even less considered features across modalities. In our view, this is a missed opportunity; there is valuable information to be extracted when combining features from different modalities. An example is the correlation between heart rate and activity. When the heart rate changes abruptly without activity, this can indicate an interesting segment for a model to detect heart problems. As another example, 1 study used timestamps as features that can provide information about seasonality in longitudinal data. This could be used to inspect (change in) circadian rhythm as a biomarker for cardiovascular disease. Interestingly, ECG morphology is well researched and used as a feature. However, no analogous decomposition of PPG signals is used in the studies. Therefore, we advocate a similar exploration of the PPG signals. Finally, we argue that in addition to the technical shortcomings discussed, societal factors (under the umbrella term ethical or socially responsible AI) must also be addressed [74]. From the patients' point of view, there are concerns regarding reliability, privacy, and especially fairness and AI bias of the system [75]. Our findings of the lack of realistic data and the imbalance in data link to the latter, as it introduces sampling bias [76], for example. A study by Parikh et al [77] refers to this as a statistical bias and argues that, especially in the medical field, there can also be social biases that are caused by inequity of

patients' access to health care (technology) or a combination of both, for example, missing data in certain subgroups. Efforts should be made to remove bias in data (before exposing to an AI model) [74] and in the model itself. This referred to as debiasing [74][76][78].

From the physicians' point of view, the performance of machine learning models is potentially reaching that of health care professionals' point of view [79][80], which brings techno-dystopic fear of rivalry between AI and human experts. The study by Di Ieva [81] offers an alternative view by stating that this fear can be overcome by considering the success of multidisciplinary teams in modern medicine and that in line with that paradigm, AI is an assisting expert in that team, rather than a competitor.

As a final note, we would like to emphasize that we did not fully perform a quality assessment of the risk of bias in the clinical data acquisition of the studies. Instead, we used the TRL to capture these risks from a machine learning perspective and describe these limitations throughout. To this end, studies with low methodological quality did not achieve a higher TRL. In addition, we did not consider conference papers as journal papers are more comprehensive and elaborate in general. However, in the field of machine learning, conferences are used to publish completed research (not limited to an abstract as in other fields). Therefore, we might have missed new developments from conference papers that have been described in detail, yet not fully scrutinized as in journal papers.

REFERENCES

[1] A. Rizwan, A. Zoha, I. B. Mabrouk, H. M. Sabbour, A. S. Al-Sumaiti, A. Alomainy, M. A. Imran, and Q. H. Abbasi. “A review on the state of the art in atrial fibrillation detection enabled by machine learning”. In: *IEEE reviews in biomedical engineering* 14 (2020), pp. 219–239.

[2] T. Pereira, N. Tran, K. Gadhoumi, M. M. Pelter, D. H. Do, R. J. Lee, R. Colorado, K. Meisel, and X. Hu. “Photoplethysmography based atrial fibrillation detection: a review”. In: *NPJ digital medicine* 3.1 (2020), p. 3.

[3] S.-K. Chua, L.-C. Chen, L.-M. Lien, H.-M. Lo, Z.-Y. Liao, S.-P. Chao, C.-Y. Chuang, and C.-Z. Chiu. “Comparison of arrhythmia detection by 24-hour Holter and 14-day continuous electrocardiography patch monitoring”. In: *Acta Cardiologica Sinica* 36.3 (2020), p. 251.

[4] S. Pradhan, J. A. Robinson, J. K. Shivapour, and C. S. Snyder. “Ambulatory arrhythmia detection with ZIO® XT patch in pediatric patients: a comparison of devices”. In: *Pediatric cardiology* 40 (2019), pp. 921–924.

[5] D. H. Brahmbhatt and M. R. Cowie. “Remote management of heart failure: an overview of telemonitoring technologies”. In: *Cardiac failure review* 5.2 (2019), p. 86.

[6] A. Van Stipdonk, S. Vanbelle, I. Ter Horst, J. Luermans, M. Meine, A. Maass, A. Auricchio, F. Prinzen, and K. Vernooy. “Large variability in clinical judgement and definitions of left bundle branch block to identify candidates for cardiac resynchronisation therapy”. In: *International journal of cardiology* 286 (2019), pp. 61–65.

[7] P. J. Slomka, D. Dey, A. Sitek, M. Motwani, D. S. Berman, and G. Germano. “Cardiac imaging: working towards fully-automated machine analysis & interpretation”. In: *Expert review of medical devices* 14.3 (2017), pp. 197–212.

[8] M. T. Trockel, N. K. Menon, S. G. Rowe, M. T. Stewart, R. Smith, M. Lu, P. K. Kim, M. A. Quinn, E. Lawrence, D. Marchalik, *et al.* “Assessment of physician sleep and wellness, burnout, and clinically significant medical errors”. In: *JAMA network open* 3.12 (2020), e2028111.

[9] C. Krittanawong, A. J. Rogers, K. W. Johnson, Z. Wang, M. P. Turakhia, J. L. Halperin, and S. M. Narayan. “Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management”. In: *Nature Reviews Cardiology* 18.2 (2021), pp. 75–91.

[10] K. Bayoumy, M. Gaber, A. Elshafeey, O. Mhaimeed, E. H. Dineen, F. A. Marvel, S. S. Martin, E. D. Muse, M. P. Turakhia, K. G. Tarakji, *et al.* “Smart wearable devices in cardiovascular care: where we are and how to move forward”. In: *Nature Reviews Cardiology* 18.8 (2021), pp. 581–599.

[11] G. D. Giebel and C. Gissel. “Accuracy of mHealth devices for atrial fibrillation screening: systematic review”. In: *JMIR mHealth and uHealth* 7.6 (2019), e13641.

[12] M. Héder. “From NASA to EU: the evolution of the TRL scale in Public Sector Innovation”. In: *The Innovation Journal* 22.2 (2017), pp. 1–23.

[13] M. Komorowski. “Artificial intelligence in intensive care: are we there yet?” In: *Intensive care medicine* 45.9 (2019), pp. 1298–1300.

[14] L. M. Fleuren, P. Thoral, D. Shillan, A. Ercole, P. W. Elbers, and R. D. R. N. C. H. M. G. B. K. T. L. G. T. R. L. F. S. E. L. G. A. RJ. “Machine learning in intensive care medicine: ready for take-off?” In: *Intensive care medicine* 46.7 (2020), pp. 1486–1488.

[15] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.* “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”. In: *bmj* 372 (2021).

[16] J. Stehlik, C. Schmalfuss, B. Bozkurt, J. Nativi-Nicolau, P. Wohlfahrt, S. Wegerich, K. Rose, R. Ray, R. Schofield, A. Deswal, *et al.* “Continuous wearable monitoring analytics predict heart failure hospitalization: the LINK-HF multicenter study”. In: *Circulation: Heart Failure* 13.3 (2020), e006513.

[17] J. Torres-Soto and E. A. Ashley. “Multi-task deep learning for cardiac rhythm detection in wearable devices”. In: *NPJ digital medicine* 3.1 (2020), p. 116.

[18] S. K. Bashar, D. Han, F. Zieneddin, E. Ding, T. P. Fitzgibbons, A. J. Walkey, D. D. McManus, B. Javidi, and K. H. Chon. “Novel density Poincaré plot based machine learning method to detect atrial fibrillation from premature atrial/ventricular contractions”. In: *IEEE Transactions on Biomedical Engineering* 68.2 (2020), pp. 448–460.

[19] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone, *et al.* “Passive detection of atrial fibrillation using a commercially available smartwatch”. In: *JAMA cardiology* 3.5 (2018), pp. 409–416.

[20] J. Wasserlauf, C. You, R. Patel, A. Valys, D. Albert, and R. Passman. “Smartwatch performance for the detection and quantification of atrial fibrillation”. In: *Circulation: Arrhythmia and Electrophysiology* 12.6 (2019), e006834.

[21] A. J. A. Majumder, Y. A. ElSaadany, R. Young Jr, and D. R. Ucci. “An energy efficient wearable smart IoT system to predict cardiac arrest”. In: *Advances in Human-Computer Interaction* 2019.1 (2019), p. 1507465.

[22] K. Lee, S. Kim, H. O. Choi, J. Lee, and Y. Nam. “Analyzing electrocardiogram signals obtained from a nymi band to detect atrial fibrillation”. In: *Multimedia Tools and Applications* 79 (2020), pp. 15985–15999.

[23] E. M. Green, R. van Mourik, C. Wolfus, S. B. Heitner, O. Dur, and M. J. Semigran. “Machine learning detection of obstructive hypertrophic cardiomyopathy using a wearable biosensor”. In: *NPJ digital medicine* 2.1 (2019), p. 57.

[24] T. Pereira, C. Ding, K. Gadhoumi, N. Tran, R. A. Colorado, K. Meisel, and X. Hu. “Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation”. In: *Physiological measurement* 40.12 (2019), p. 125002.

[25] S. Kwon, J. Hong, E.-K. Choi, B. Lee, C. Baik, E. Lee, E.-R. Jeong, B.-K. Koo, S. Oh, and Y. Yi. “Detection of atrial fibrillation using a ring-type wearable device (CardioTracker) and deep learning analysis of photoplethysmography signals: prospective observational proof-of-concept study”. In: *Journal of Medical Internet Research* 22.5 (2020), e16443.

[26] S. Kwon, J. Hong, E.-K. Choi, E. Lee, D. E. Hostallero, W. J. Kang, B. Lee, E.-R. Jeong, B.-K. Koo, S. Oh, *et al.* “Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study”. In: *JMIR mHealth and uHealth* 7.6 (2019), e12770.

[27] V. D. Corino, R. Laureanti, L. Ferranti, G. Scarpini, F. Lombardi, and L. T. Mainardi. “Detection of atrial fibrillation episodes using a wristband device”. In: *Physiological measurement* 38.5 (2017), p. 787.

[28] A. J. Shah, N. Isakadze, O. Levantsevych, A. Vest, G. Clifford, and S. Nemati. “Detecting heart failure using wearables: a pilot study”. In: *Physiological measurement* 41.4 (2020), p. 044001.

[29] E. Chen, J. Jiang, R. Su, M. Gao, S. Zhu, J. Zhou, and Y. Huo. “A new smart wristband equipped with an artificial intelligence algorithm to detect atrial fibrillation”. In: *Heart rhythm* 17.5 (2020), pp. 847–853.

[30] G. Quer, P. Gouda, M. Galarnyk, E. J. Topol, and S. R. Steinhubl. “Inter-and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults”. In: *Plos one* 15.2 (2020), e0227709.

[31] E. Jeon, K. Oh, S. Kwon, H. Son, Y. Yun, E.-S. Jung, M. S. Kim, *et al.* “A lightweight deep learning model for fast electrocardiographic beats classification with a wearable cardiac monitor: development and validation study”. In: *JMIR Medical Informatics* 8.3 (2020), e17037.

[32] D. Lai, Y. Bu, Y. Su, X. Zhang, and C.-S. Ma. “A flexible multilayered dry electrode and assembly to single-lead ECG patch to monitor atrial fibrillation in a real-life scenario”. In: *IEEE Sensors Journal* 20.20 (2020), pp. 12295–12306.

[33] M. Altini, P. Casale, J. Penders, and O. Amft. “Cardiorespiratory fitness estimation in free-living using wearable sensors”. In: *Artificial intelligence in medicine* 68 (2016), pp. 37–46.

[34] F. P. Akbulut and A. Akan. “A smart wearable system for short-term cardiovascular risk assessment with emotional dynamics”. In: *Measurement* 128 (2018), pp. 237–246.

[35] L. J. Mena, V. G. Félix, A. Ochoa, R. Ostos, E. González, J. Aspuru, P. Velarde, and G. E. Maestre. “Mobile personal health monitoring for automated classification of electrocardiogram signals in elderly”. In: *Computational and mathematical methods in medicine* 2018.1 (2018), p. 9128054.

[36] Y. Xia and Y. Xie. “A novel wearable electrocardiogram classification system using convolutional neural networks and active learning”. In: *Ieee Access* 7 (2019), pp. 7989–8001.

[37] F. Miao, X. Wang, L. Yin, and Y. Li. “A wearable sensor for arterial stiffness monitoring based on machine learning algorithms”. In: *IEEE Sensors Journal* 19.4 (2018), pp. 1426–1434.

[38] Y.-S. Kim, M. Mahmood, Y. Lee, N. K. Kim, S. Kwon, R. Herbert, D. Kim, H. C. Cho, and W.-H. Yeo. “All-in-one, wireless, stretchable hybrid electronics for smart, connected, and ambulatory physiological monitoring”. In: *Advanced Science* 6.17 (2019), p. 1900939.

[39] A. Sharma, N. Garg, S. Patidar, R. San Tan, and U. R. Acharya. “Automated pre-screening of arrhythmia using hybrid combination of Fourier–Bessel expansion and LSTM”. In: *Computers in biology and medicine* 120 (2020), p. 103753.

[40] J. J. Oresko, Z. Jin, J. Cheng, S. Huang, Y. Sun, H. Duschl, and A. C. Cheng. “A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing”. In: *IEEE Transactions on Information Technology in Biomedicine* 14.3 (2010), pp. 734–740.

[41] M. Sadrawi, C.-H. Lin, Y.-T. Lin, Y. Hsieh, C.-C. Kuo, J. C. Chien, K. Haraikawa, M. F. Abbod, and J.-S. Shieh. “Arrhythmia evaluation in wearable ECG devices”. In: *Sensors* 17.11 (2017), p. 2445.

[42] R. Allami. “Premature ventricular contraction analysis for real-time patient monitoring”. In: *Biomedical Signal Processing and Control* 47 (2019), pp. 358–365.

[43] Y. Zhao, Z. Shang, and Y. Lian. “A 13.34 μW event-driven patient-specific ANN cardiac arrhythmia classifier for wearable ECG sensors”. In: *IEEE transactions on biomedical circuits and systems* 14.2 (2019), pp. 186–197.

[44] J. Wu, F. Li, Z. Chen, Y. Pu, and M. Zhan. “A neural network-based ECG classification processor with exploitation of heartbeat similarity”. In: *IEEE Access* 7 (2019), pp. 172774–172782.

[45] S. Saadatnejad, M. Oveisi, and M. Hashemi. “LSTM-based ECG classification for continuous monitoring on personal wearable devices”. In: *IEEE journal of biomedical and health informatics* 24.2 (2019), pp. 515–523.

[46] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza. “Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems”. In: *IEEE transactions on biomedical circuits and systems* 12.5 (2018), pp. 982–992.

[47] D. Dubin. *Rapid interpretation of EKG’s: an interactive course*. Cover Publishing Company, 2000.

[48] X. Tang, Z. Ma, Q. Hu, and W. Tang. “A real-time arrhythmia heartbeats classification algorithm using parallel delta modulations and rotated linear-kernel support vector machines”. In: *IEEE Transactions on Biomedical Engineering* 67.4 (2019), pp. 978–986.

[49] K. Feng, X. Pi, H. Liu, and K. Sun. “Myocardial infarction classification based on convolutional neural network and recurrent neural network”. In: *Applied Sciences* 9.9 (2019), p. 1879.

[50] H. Zhang, Z. Dong, J. Gao, P. Lu, and Z. Wang. “Automatic screening method for atrial fibrillation based on lossy compression of the electrocardiogram signal”. In: *Physiological measurement* 41.7 (2020), p. 075005.

[51] S.-F. Huang and H.-P. Lu. “Classification of temporal data using dynamic time warping and compressed learning”. In: *Biomedical Signal Processing and Control* 57 (2020), p. 101781.

[52] W. Chou, P.-J. Wu, C.-C. Fang, Y.-S. Yen, and B.-S. Lin. “Design of smart brain oxygenation monitoring system for estimating cardiovascular disease severity”. In: *IEEE Access* 8 (2020), pp. 98422–98429.

[53] C. Yang, N. D. Aranoff, P. Green, and N. Tavassolian. “Classification of aortic stenosis using time–frequency features from chest cardio-mechanical signals”. In: *IEEE Transactions on Biomedical Engineering* 67.6 (2019), pp. 1672–1683.

[54] A. Amirshahi and M. Hashemi. “ECG classification algorithm based on STDP and R-STDP neural networks for real-time monitoring on ultra low-power personal wearable devices”. In: *IEEE transactions on biomedical circuits and systems* 13.6 (2019), pp. 1483–1493.

[55] Z. Yan, J. Zhou, and W.-F. Wong. “Energy efficient ECG classification with spiking neural network”. In: *Biomedical Signal Processing and Control* 63 (2021), p. 102170.

[56] H. W. Lui and K. L. Chow. “Multiclass classification of myocardial infarction with convolutional and recurrent neural networks for portable ECG devices”. In: *Informatics in Medicine Unlocked* 13 (2018), pp. 26–33.

[57] M. Shao, Z. Zhou, G. Bin, Y. Bai, and S. Wu. “A wearable electrocardiogram telemonitoring system for atrial fibrillation detection”. In: *Sensors* 20.3 (2020), p. 606.

[58] M. Gilani, J. M. Eklund, and M. Makrehchi. “Automated detection of atrial fibrillation episode using novel heart rate variability features”. In: *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2016, pp. 3461–3464.

[59] Z. Mei, X. Gu, H. Chen, and W. Chen. “Automatic atrial fibrillation detection based on heart rate variability and spectral features”. In: *IEEE Access* 6 (2018), pp. 53566–53575.

[60] D. S. Quintana. “Statistical considerations for reporting and planning heart rate variability case-control studies”. In: *Psychophysiology* 54.3 (2017), pp. 344–349.

[61] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[62] O. T. Inan, M. Baran Pouyan, A. Q. Javaid, S. Dowling, M. Etemadi, A. Dorier, J. A. Heller, A. O. Bicen, S. Roy, T. De Marco, *et al.* “Novel wearable seismocardiography and machine learning algorithms can assess clinical status of heart failure patients”. In: *Circulation: Heart Failure* 11.1 (2018), e004313.

[63] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, *et al.* “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure”. In: *Ecography* 40.8 (2017), pp. 913–929.

[64] Q.-a. Mastoi, T. Y. Wah, and R. Gopal Raj. “Reservoir computing based echo state networks for ventricular heart beat classification”. In: *Applied Sciences* 9.4 (2019), p. 702.

[65] F. Ma, J. Zhang, W. Liang, and J. Xue. “Automated classification of atrial fibrillation using artificial neural network for wearable devices”. In: *Mathematical Problems in Engineering* 2020.1 (2020), p. 9159158.

[66] N. Wang, J. Zhou, G. Dai, J. Huang, and Y. Xie. “Energy-efficient intelligent ECG monitoring for wearable devices”. In: *IEEE transactions on biomedical circuits and systems* 13.5 (2019), pp. 1112–1121.

[67] A. Scirè, F. Tropeano, A. Anagnostopoulos, and I. Chatzigiannakis. “Fog-computing-based heartbeat detection and arrhythmia classification using machine learning”. In: *Algorithms* 12.2 (2019), p. 32.

[68] X. Fan, Q. Yao, Y. Cai, F. Miao, F. Sun, and Y. Li. “Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings”. In: *IEEE journal of biomedical and health informatics* 22.6 (2018), pp. 1744–1753.

[69] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[70] S. Kiranyaz, T. Ince, and M. Gabbouj. “Real-time patient-specific ECG classification by 1-D convolutional neural networks”. In: *IEEE transactions on biomedical engineering* 63.3 (2015), pp. 664–675.

[71] G. B. Moody and R. G. Mark. “The impact of the MIT-BIH arrhythmia database”. In: *IEEE engineering in medicine and biology magazine* 20.3 (2001), pp. 45–50.

[72] I. Balki, A. Amirabadi, J. Levman, A. L. Martel, Z. Emersic, B. Meden, A. Garcia-Pedrero, S. C. Ramirez, D. Kong, A. R. Moody, *et al.* “Sample-size determination methodologies for machine learning in medical imaging research: a systematic review”. In: *Canadian Association of Radiologists Journal* 70.4 (2019), pp. 344–353.

[73] A. N. Richter and T. M. Khoshgoftaar. “Sample size determination for biomedical big data with limited labels”. In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 9 (2020), pp. 1–13.

[74] L. Cheng, K. R. Varshney, and H. Liu. “Socially responsible ai algorithms: Issues, purposes, and challenges”. In: *Journal of Artificial Intelligence Research* 71 (2021), pp. 1137–1181.

[75] V.-T. Tran, C. Riveros, and P. Ravaud. “Patients’ views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort”. In: *NPJ digital medicine* 2.1 (2019), p. 53.

[76] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim. “Mitigating bias in machine learning for medicine”. In: *Communications medicine* 1.1 (2021), p. 25.

[77] R. B. Parikh, S. Teeple, and A. S. Navathe. “Addressing bias in artificial intelligence in health care”. In: *Jama* 322.24 (2019), pp. 2377–2378.

[78] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He. “Bias and debias in recommender system: A survey and future directions”. In: *ACM Transactions on Information Systems* 41.3 (2023), pp. 1–39.

[79] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, *et al.* “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis”. In: *The lancet digital health* 1.6 (2019), e271–e297.

[80] W. P. van Doorn, P. M. Stassen, H. F. Borggreve, M. J. Schalkwijk, J. Stoffers, O. Bekers, and S. J. Meex. “A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis”. In: *PLoS One* 16.1 (2021), e0245157.

[81] A. Di Ieva. “AI-augmented multidisciplinary teams: hype or hope?” In: *The Lancet* 394.10211 (2019), p. 1801.

3

DATA-EFFICIENT ML METHODS IN THE ME-TIME STUDY: RATIONALE AND DESIGN OF A LONGITUDINAL STUDY TO DETECT ATRIAL FIBRILLATION AND HEART FAILURE FROM WEARABLES

Smartwatches enable continuous and non-invasive timeseries monitoring of cardiovascular biomarkers like heart rate (from photoplethysmograms), step counter, skin temperature, etcetera, as such they have a promise in assisting in early detection and prevention of cardiovascular disease. Although these biomarkers may not be directly useful to physicians, a machine learning (ML) model could find clinically relevant patterns. Unfortunately, machine learning models typically need supervised, i.e. annotated, data, and labeling of large amounts of continuous data is very labour intensive. Therefore, ML methods that are data-efficient, i.e. needing a low number of labels, are required to detect potential clinical value in patterns found in wearable data.
The primary study objective of the ME-TIME (Machine learning Enabled TIme series analysis in MEdicine) study is to design a ML model that

This chapter has been published in Cardiovascular Digital Health **4**, 6 (2023).

can detect atrial fibrillation (AF) and heart failure (HF) from wearable data in a data-efficient manner. To achieve this, self-supervised and weakly-supervised learning techniques are used. Two hundred subjects (100 reference, 50 AF and 50 HF) are being invited to participate in wearing a Fitbit fitness tracker for three months.
Enrollment began May 1st. Interested volunteers are sent a questionnaire to determine their health, in particular cardiovascular health. Volunteers without any (history of) serious illness are assigned to the reference group. Participants with AF and HF are recruited in the Haga teaching hospital in The Hague, The Netherlands. Using self-supervised and multiple instance learning techniques we hypothesize that patterns specific to AF and HF can be found in continuous data obtained from smartwatches.

3.1. INTRODUCTION

Cardiovascular disease is one of the leading causes of mortality globally[1] and cardiovascular healthcare accounts for a large portion of global healthcare costs and expenses. Early detection and prevention will decrease the burden of cardiovascular disease and will therefore decrease mortality, morbidity, and costs. Cardiovascular monitoring using big data from wearables and machine learning can drastically increase the availability and efficiency of cardiovascular healthcare globally, at the fraction of the costs of conventional medical-grade devices.

Electrocardiogram (ECG) monitoring, such as Holters or implantable loop recorders, are the gold standard for monitoring of outpatients with known or suspected arrhythmias. However, they are burdensome, can only be used for a limited period of time, and are expensive. Implantable loop recorders are invasive and have to be manually activated and analyzed in the hospital. This severely limits the use of these devices for long-term home monitoring of patients, and they have suboptimal patient comfort. For patients with chronic cardiovascular disease, such as atrial fibrillation and heart failure, this implies frequent hospital visits and sometimes even hospital admissions (associated with higher mortality) that can be prevented by continuous and adequate home monitoring.

With the widespread availability of reliable, consumer-grade wearables such as smartwatches, continuous monitoring of, for example, heart rate with photoplethysmography and step counting with accelerometers is possible. This monitoring is easy, patient friendly, and cost effective. Combining the power of large amounts of data (big data) and novel machine learning techniques, these time series can be used to detect and perhaps even predict cardiovascular disease, therefore improving patient care. There are some caveats, however, as not all wearables have the same characteristics and quality. Consequently, they have been used with moderate success[2][3]. They also provide less informative diagnostic signals as compared to, for example, electrocardiography or other commonly used cardiologic diagnostic modalities. The challenge but also the strength of machine learning models is that they learn by example and therefore large amounts of data are needed for which the cardiovascular outcome (class label) has been determined. Typically, supervised learning is used, where each observation of the data has a class label. This must be done with ECGs, since photoplethysmography or derived signals are difficult to interpret by a clinician. This so-called labeling or annotating of signals by physicians is infeasible for the large amounts of (continuous) data required, and therefore semi-automated[4][5] and fully automated[2][3] ECG labeling systems[6] have been developed. However, these still require a lot of manual labor from continuously monitored users.

Therefore, the objective of the ME-TIME is (early) detection and prevention of heart disease by leveraging time series data from smartwatches, a cloud-based infrastructure, and machine learning algorithms specifically designed to function effectively with minimal labeling efforts.

3.2. MATERIALS AND METHODS

3.2.1. STUDY DESIGN & DATA COLLECTION

ME-TIME (registered at ClinicalTrials.gov; ID: NCT05802563) is designed as an observational cohort study consisting of 3 data subject groups, as depicted in Figure 3.1. The first group consists of patients with systolic heart failure (HF group); the second group consists of patients with documented atrial fibrillation (AF group); and the third group, serving as a reference, consists of healthy volunteers. The rationale for creating distinct AF and HF groups comes from their unique pathophysiological characteristics. Consequently, heart rate patterns that are indicative of these diseases might also be different. The HF group consists of 50 study participants with systolic heart failure, defined as a left ventricular ejection fraction <35% without documented atrial fibrillation. The AF group consists of 50 patients with documented atrial fibrillation (paroxysmal, persistent, or permanent) without systolic heart failure. Ejection fractions will be assessed from echocardiograms that are made within 1 year of inclusion, and if this is not available an echocardiogram will be performed. The reference group consists of 100 participants without any prior medical history and without medication use. Potential study subjects that meet any of the following criteria will be excluded from participation in this study: age <18 years, age >85 years, recent pulmonary venous antrum isolation (<1 year), kidney or liver failure, known systemic active inflammatory disease, impaired mental state, inability to use a fitness tracker or mobile phone, impaired cognition, and inability to understand the study protocol. Patients will be asked by their treating physician if they may be approached by an investigator to inform them about the study and potential participation. Healthy participants are recruited through local advertising. Anyone that is interested will then receive an information brochure and informed consent form. At least 2 days after the patient's receipt of the brochure, the research team will call the patient to schedule an appointment. During this visit, the patient submits the signed consent form and will undergo an ECG and blood pressure measurement that will be analyzed by an experienced cardiologist (I.B.). Participants can use their own Fitbit and are otherwise provided with a Fitbit Inspire 2 or Fitbit Charge 5 smartwatch. The device type is assigned to a participant at random to prevent device sampling bias. This will also help to investigate the effect of device type on the performance of the final model. A Fitbit

Figure 3.1.: Data analysis pipeline for the ME-TIME study. Included participants (image 1) are given a smartwatch (image 2), which is connected to our data acquisition and storage platform (image 3). The resulting data are then preprocessed (image 4) and put into the data-efficient machine learning model (image 5). AF = atrial fibrillation group; HF = heart failure group; Ref = reference group.

account will be created for all participants which will be connected to a custom-built data platform using the Google Cloud Platform. Our platform features a data portal for research staff to easily register or deregister participants by authorizing a connection to their Fitbit data. Data are extracted daily from Fitbits until the observation period ends and can be analyzed either in the cloud or locally.
All participants will be asked to fill out a survey regarding their health. All participants are monitored for a period of 3 months. After written consent from the 200 subjects, heart rate, step counter, and sleep time series data are extracted from the data platform. Clinical metadata such as age, height, weight, blood pressure at baseline, health survey, and medication use are saved in the Castor (Ciwit BV, Amsterdam, The Netherlands) electronic database.

3.2.2. DATA PRIVACY

After performing a thorough data protection impact assessment, the local hospital security information and privacy officers granted permission to perform the study. This was also validated by the ethics review board. The data protection impact assessment describes a data management plan conforming to the European General Data Protection Regulation. To protect the data privacy of the participants, all data are pseudonymized. Only the researchers have access to the sensor data, and only the Principal Investigator has access to personal information of participants (ie, names, contact information, etc). They have all signed processing agreements. Second, the Google servers storing the data are only located within the Netherlands; hence the data does not leave the country, therefore conforming to Dutch law. This is done to have a clear data infrastructure both legally and technically to explain to participants.

3.2.3. DATA CHARACTERISTIC & PREPARATION

The data first undergoes a process involving resampling and artifact removal. In our experience with Fitbit smartwatches, the heart rate is nonuniformly sampled, with a prevalent rate of 0.2 Hz. Therefore, the heart rate is resampled to once per 5 seconds. The step counter is sampled once per minute.

Artifacts involving samples with numerous consecutive constant values are removed, if more than 12 consecutive constant values (equivalent to 1 minute of heart rate samples) are detected. This 1-minute threshold was chosen based on visual inspection, which revealed that heart rate patterns typically occur in the order of minutes, often spanning 10–20 minutes. For sequences with fewer than 12 consecutive missing values, linear interpolation is applied. From the cleaned time series, smaller segments, denoted as windows, are extracted and employed as input for a machine learning model. This process involves a sliding window and windows containing time gaps are excluded. Windows have 2 design considerations: the window size, which determines the number of samples within a window and defines its dimensionality, and the stride, which establishes the step size dictating the shift between windows.

Although the cardiovascular condition of each subject is known, it is unknown in which specific windows these conditions manifest themselves. This is owing to the paroxysmal nature of atrial fibrillation and the variable symptoms of heart failure, which can be influenced by factors like medication adjustments, dietary changes, and the disease's progressive course. In other words, the subject label is known, but the individual window labels are unknown. This is visually represented in Figure 3.2a, where the subject label is depicted by the blue/red colors and the unknown window labels are indicated by black dotted lines.

3.2.4. PLANNED MACHINE LEARNING APPROACH

Our planned machine learning approach is tailored to operate in this setting through a 2-stage process. To learn informative patterns/features directly from the input data, despite the lack of labeled windows, the first stage involves using self-supervised learning. A commonly used self-supervised learning technique involves compressing the input windows to a lower-dimensional representation and then reconstructing the original input from this compact representation, as depicted in Figure 3.2b [7][8] Instead of reconstruction, another technique is to forecast future time points of the input data[9]. The second stage involves multiple-instance learning (MIL). MIL, depicted in Figure 3.3 and more elaborately explained in Box 1, is suitable for data where a single prediction is made collectively on a group of samples (known as a "bag") instead of predicting on individual samples (known as "instances"). MIL techniques align well with our time series data, where during the

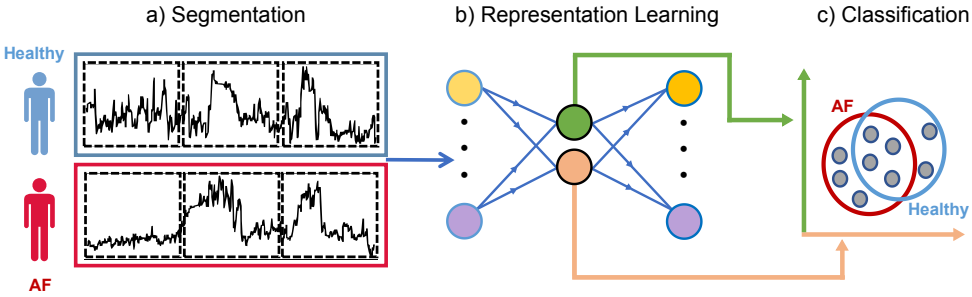


Figure 3.2.: Pipeline for the study's proposed approach. a: Segmentation of the time series of each subject (1 healthy and 2 atrial fibrillation) using a sliding window. Only the label of the entire subject is available, instead of each individual window. b: The windows are inputs to an autoencoder and are compressed to a smaller (2-dimensional for illustrative purposes) representation. c: The compressed representation is used to train a multiple-instance classifier that can distinguish between healthy, atrial fibrillation, or heart failure). AF = atrial fibrillation.

training phase only 1 label related to the subject (bag label) is known while the individual labels of the compressed windows (instance labels) remain unknown. In the testing phase the primary objective is to predict 1 clinical outcome for each subject. The key concept in MIL is that each bag of instances is labeled as positive (heart disease) if it contains a certain amount of positive instances and negative (healthy) if it contains no positive instances (only negative instances). Although traditional MIL approaches often classify a bag as positive even with just 1 positive instance, we aim to minimize false-positives by setting a threshold on the number of positive instances required to label a bag as positive. This threshold will be determined through hyperparameter tuning. Thus, instead of learning a model that predicts the cardiovascular outcome of individual instances, we are learning a model that predicts the outcome of a bag of instances.

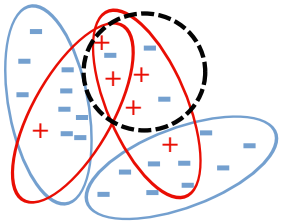


Figure 3.3.: Illustration of multiple-instance learning. The red and blue lines indicate bags of heart disease patients and reference subjects. Even though the labels for each instance are not known, for the sake of this example, the plus and minus signs depict time windows where heart disease is present or absent, respectively. The decision boundary is depicted by a dotted circle, where instances within the circle are classified as heart disease, and instances outside the circle are classified as healthy.

A simple MIL example
MIL is elaborated with an example in three steps.
Initial setup: Under traditional supervised learning, each window must be annotated to train a machine learning model. However, in our MIL setting (Figure 3.3), only the bag label is known for the entire set of windows related to a subject. A bag label is considered negative if none of the subject's individual samples are associated with heart disease, indicating that the subject is not affected by it. Conversely, a bag label is positive if a certain amount of the subject's samples is linked to heart disease.
Learning: The algorithm then learns a model based on the bag-level labels only. The goal of the MIL algorithm is to learn a model that can correctly predict the bag-level labels given the instances in each bag. By training on the bag-level labels, the MIL algorithm can capture patterns and relationships within the data that help identify the presence or absence of heart disease. Note that the decision boundary produced by the model in Figure 3.3 is not ideally suited for classifying individual windows, which is expected, as it did not use this information. However, if a sufficient number of windows are classified accurately, the correct bag label can still be predicted. This is accomplished during training by aggregating these accurate classifications using methods like majority voting, or by setting a threshold for the minimum number of positively predicted windows needed to assign a positive bag label. **Prediction:** Once the model is trained, it can predict the label of a new bag by examining the instances in the bag. If the model predicts that a certain percentage of instances in the bag are positive defined by the threshold, the bag is classified as positive (heart disease) and negative (healthy) otherwise. By analyzing the presence or absence of positive instances within the bag, the MIL algorithm can make predictions on a bag level, providing insights into the subject's condition.

3.2.5. ALGORITHM VALIDATION

In our specific setting where the model encounters data from previously unseen subjects without any prior knowledge, we use leave-p-subjects-out cross-validation (LPSOCV). This approach, shown in Figure 3.4, ensures a more accurate reflection of real-world situations. LPSOCV involves multiple iterations, or folds, during which data from distinct subjects are used for training and validation purposes (ie, the model is validated on data from subjects that the model is not trained on), mitigating observation bias. Machine learning models are sensitive to

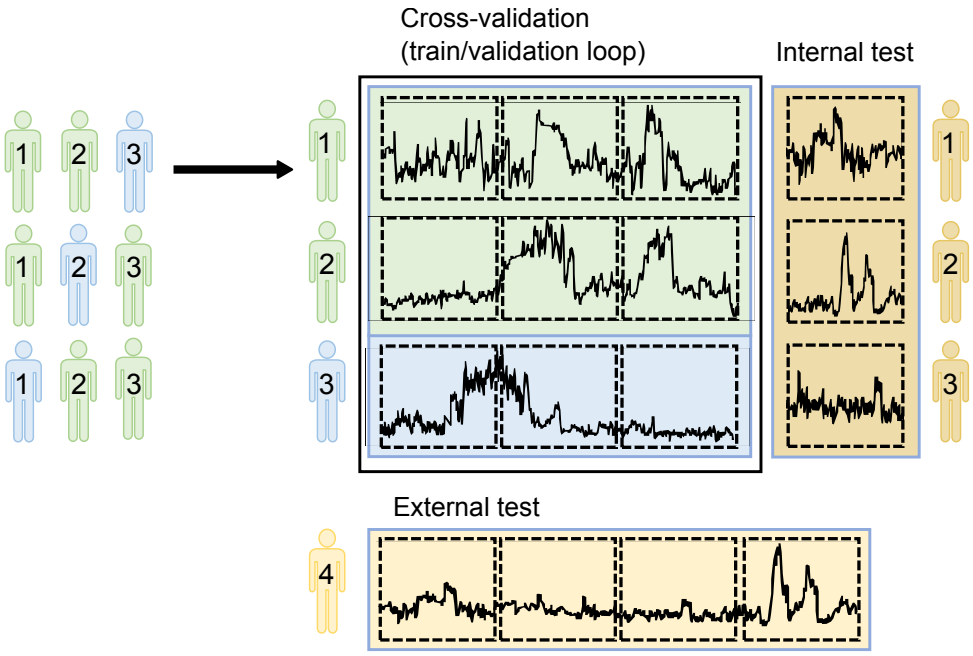


Figure 3.4.: Our leave-p-subjects-out cross-validation strategy consists of the following: On the left, cross-validation folds are illustrated using 3 subjects with corresponding subject number. Green and blue represent training and validation subjects, respectively. On the right, a single fold is expanded and additionally illustrates the internal and external test sets. Each row corresponds to a subject, and the dotted squares within each row represent windows. The yellow external test block encompasses entire subjects and their corresponding data points that have not been encountered by the model. Meanwhile, the dark yellow internal test block consists of unobserved data points, representing the final 20% measurements of the time series from subjects already encountered during model development.

the distribution of classes. To mitigate potential bias owing to different class distributions in each fold, we incorporate stratification during the cross-validation process. This ensures that the ratio of nonarrhythmic, atrial fibrillation, and heart failure subjects remains approximately consistent and that the influence of inconsistent class distribution across different folds is minimized.
However, Fitbit time series data exhibit inter-subject variability resulting from individuals' distinct physical attributes[10], making the development of a universally effective model for "new" subjects challenging. In order to examine the effect of inter-subject variability, we will assess the model on 2 distinct test sets. The first is an external test set that consists of subjects not previously encountered, randomly selected to make up 20% of the total subjects, with an equal number from each class. This allows us to evaluate the model's generalization capabilities. The second test set is an internal one, encompassing the final 20% of data from subjects previously encountered by the model. This segment of data was excluded during the cross-validation phase and serves as a baseline, as it minimizes the influence of inter-subject variability, providing a reliable reference for comparison.
Parameters not directly learned by the machine learning model, such as window parameters, are termed hyperparameters. Since optimal values are typically unknown in advance, multiple options are examined during LPSOCV, and the best-performing one, with the best average performance over all folds, is chosen for the final model; a process known as hyperparameter tuning.

3.3. RESULTS

Preliminary findings are discussed in the following sections.

3.3.1. PRELIMINARY FINDINGS

So far, 62 of the 200 envisioned subjects have been included and data from 22 subjects (15 healthy, 7 AF) have been extracted successfully from the data platform for preliminary analysis (table 3.1).

3.3.2. DATA SHOWS LARGE INTER-SUBJECT VARIABILITY

Non-overlapping 1-hour windows are used to segment heart rate and step counter time series data from 6 subjects. To visualize this high-dimensional data, UMAP[11] is employed to reduce the data to 2 dimensions while maintaining as much structure as possible. The resulting embedding is displayed in Figure 3.5, where the distribution of the 2-dimensional UMAP samples are illustrated per subject. When there is little overlap between subjects, finding a shared pattern among

Table 3.1.: Characteristics of preliminary study participants as of May 2022. AF = atrial fibrillation group; BMI = body mass index; HF = heart failure group; Ref = reference group.

Characteristic		Ref	HF	AF
Total		25	15	22
Age	18–39	16	1	0
	40–54	3	3	2
	55–64	5	5	3
	65+	1	6	17
Sex	Male	12	22	15
	Female	13	3	7
BMI	18.5–24.9	14	3	7
	25–29.9	6	6	8
	30+	5	6	7
Device	Charge 5	23	8	12
	Inspire 2	2	7	10

them becomes challenging, making it difficult for a model to learn. As a result, the performance of a machine learning model could be impacted as, during testing, a subject can significantly deviate from the subjects on which the model was trained.

3.3.3. Heart rate peak alignment in acceleration-deceleration curves indicate difference between 7 AF patients and 15 healthy controls

Next, we explored the heart rate recovery curves after activity (acceleration-deceleration curves)[12]. First a peak is detected, whereafter the start (onset) and end (recovery) points associated to that peak are determined by the minimum heart rate value 5 minutes before and 15 minutes after the peak. The curves are preprocessed by aligning the peaks on the time axis. Additionally, for every subject, the amplitude of the curves is rescaled by the average peak value across all curves for that individual. Figure 3.6 shows the curves for light activity, defined by a maximum of 20 steps in the 5 minutes preceding the peak and fewer than 10 steps in the 15 minutes after the peak. There are 2 noticeable differences in heart rate patterns between persistent AF patients (in red) and healthy participants (in blue). The standard deviation for AF patients is considerably smaller than that of healthy

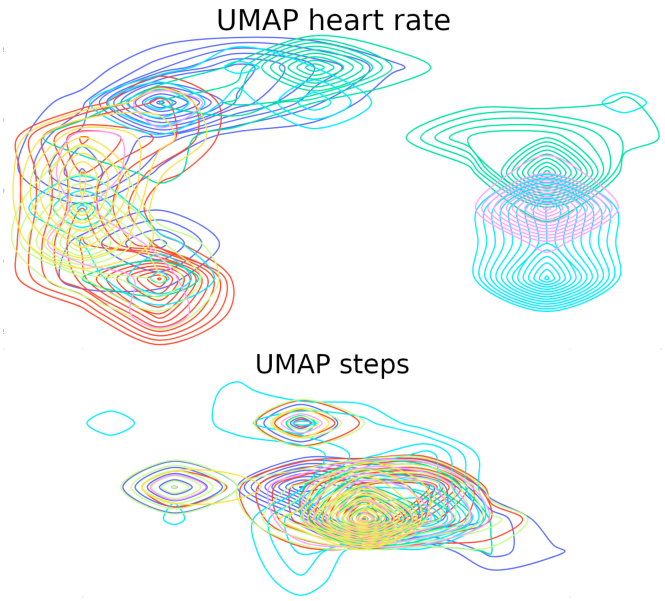


Figure 3.5.: Visualization of heart rate windows (upper image) and step windows (lower image) of 6 subjects. Each window has data of 1 hour (720 time points for heart rate, 60 for steps, respectively). Each high-dimensional window is mapped to a 2-dimensional (2D) location using UMAP.11 The contour curves illustrate the distribution of the 2D UMAP samples for each subject, with each color representing 1 of the 6 subjects.

individuals, and their heart rate recovery is slower, as observed at the 6-minute mark. These distinctions could potentially serve as clinical indicators for atrial fibrillation.

3.3.4. MIL CAN DETECT HEALTHY CARDIOVASCULAR OUTCOMES

The peak aligned acceleration-deceleration curves are concatenated with their corresponding step counter data and grouped per week to form bags. The MILES (Multiple-Instance Learning via Embedded Instance Selection) model is then used to classify every week as healthy or AF. The results in Table 3.2 show that even though the sensitivity is low, the specificity is decent. This shows potential in avoiding unnecessary visits to a cardiologist for patients who have symptoms that are wrongly suspected to be related to heart problems.

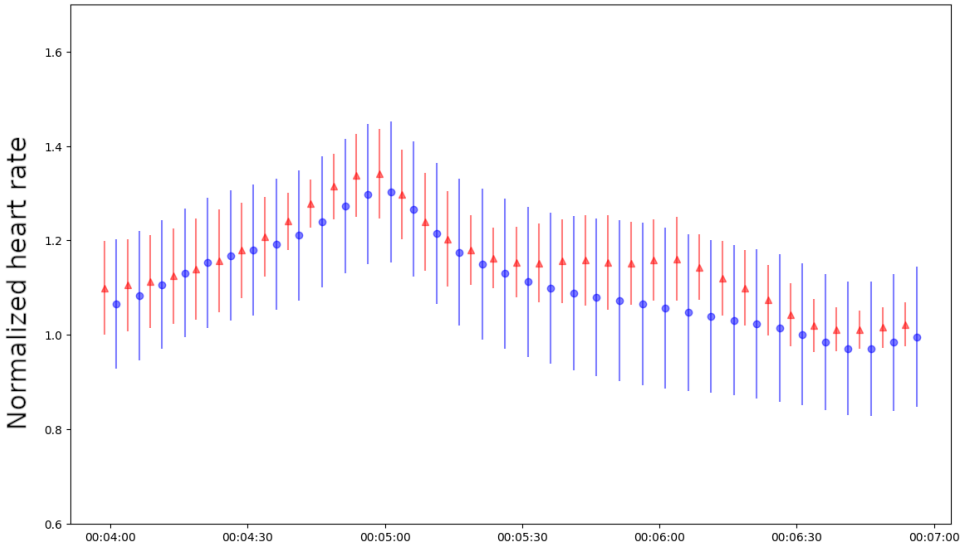


Figure 3.6.: Acceleration-deceleration curves during light activity. Red and blue represent data from subjects with persistent atrial fibrillation (n=7) and no heart disease (n=15), respectively. The mean and standard deviation are shown per time point (5 second intervals) for both groups.

Table 3.2.: Confusion matrix of per-week healthy vs atrial fibrillation classification of the MILES model with peak aligned curves concatenated with step counter data, with true and predicted labels shown vertically and horizontally, respectively.

True \ Predicted	**AF**	**Healthy**
AF	11	14
Healthy	7	33

3.3.5. STEP COUNTER AND HEART RATE ARE CORRELATED WITH A TIME DELAY

Next, we examined whether the cross-correlation function between the heart rate window and its corresponding step counter window is indicative of heart disease. To calculate the correlation, we consider varying window sizes and time differences (lags) between heart rate and steps. The computed cross-correlation matrix for the healthy group, along with the AF and HF patient groups, as shown in Figure 3.7, shows that the heart rate is correlated with the step counter with 1 minute delay.

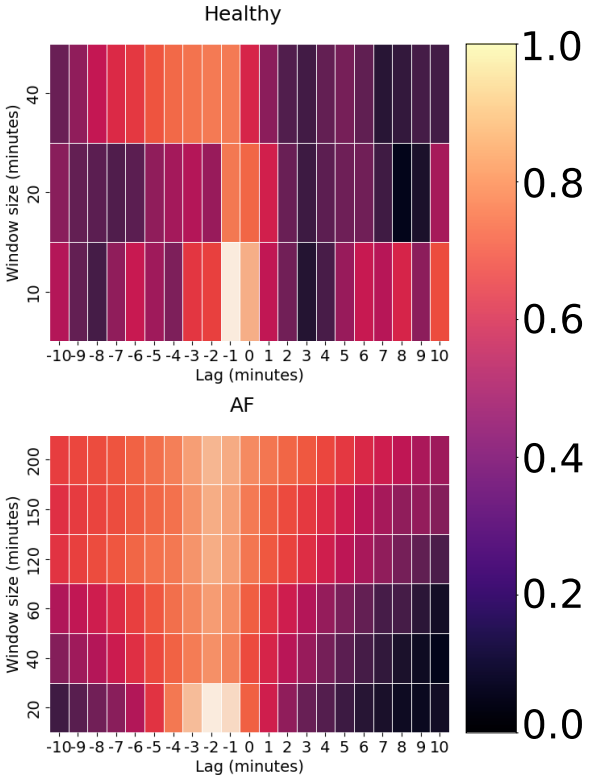


Figure 3.7.: Cross-correlation matrices between windowed heart rate data and number of steps for healthy and the persistent atrial fibrillation (AF) group. Rows are window sizes and columns lag between heart rate and step counter.

3.4. DISCUSSION

By building a suitable infrastructure with Cloud technology, big data acquired in the study is used to develop data-efficient models using methods from multiple instance and self-supervised learning.
We aim to examine the influence of inter-subject variability on predicting cardiovascular disease and will explore potential methods to mitigate these variabilities[13][14]. We expect that patterns indicative of cardiovascular disease become apparent within a timeframe of minutes, hours, or more, considering that consumer-grade wearables have a slower sampling rate compared to the gold standard. We have shown 1 example of such a pattern: the acceleration-deceleration curve. Preselecting windows based on such patterns furthermore mitigates searching through substantial amounts of data that may not provide much information about cardiovascular disease. Inspecting the cross-

correlation for several combinations of window size and lag show a different profile for healthy and AF group individuals, showing that it is meaningful to analyze step counter and heart rate together. Big data for heart disease detection requires substantial labeling efforts from physicians.
Using self-supervised learning and MIL, a model can be trained with much fewer labels. Our findings demonstrate this by employing MILES to achieve high specificity, which can aid in ruling out heart disease in individuals experiencing symptoms similar to heart disease but without the condition (ie, false-positives).

3.5. CONCLUSION

The ongoing ME-TIME study is a longitudinal observational study that uses machine learning with time series data from consumer-grade smartwatches to detect atrial fibrillation and heart failure. This will contribute to cost-effective cardiovascular monitoring of outpatients, thereby reducing exacerbation of cardiovascular disease and effectively increasing capacity of global cardiovascular healthcare.

REFERENCES

[1] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Benziger, *et al.* "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study". In: *Journal of the American College of Cardiology* 76.25 (2020), pp. 2982–3021.

[2] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone, *et al.* "Passive detection of atrial fibrillation using a commercially available smartwatch". In: *JAMA cardiology* 3.5 (2018), pp. 409–416.

[3] J. Wasserlauf, C. You, R. Patel, A. Valys, D. Albert, and R. Passman. "Smartwatch performance for the detection and quantification of atrial fibrillation". In: *Circulation: Arrhythmia and Electrophysiology* 12.6 (2019), e006834.

[4] L. Zhu, V. Nathan, J. Kuang, J. Kim, R. Avram, J. Olgin, and J. Gao. "Atrial fibrillation detection and atrial fibrillation burden estimation via wearables". In: *IEEE Journal of Biomedical and Health Informatics* 26.5 (2021), pp. 2063–2074.

[5] S. A. Lubitz, A. Z. Faranesh, C. Selvaggi, S. J. Atlas, D. D. McManus, D. E. Singer, S. Pagoto, M. V. McConnell, A. Pantelopoulos, and A. S. Foulkes. "Detection of atrial fibrillation in a large population using wearable devices: the Fitbit heart study". In: *Circulation* 146.19 (2022), pp. 1415–1424.

[6] A. Hall, A. R. J. Mitchell, L. Wood, and C. Holland. "Effectiveness of a single lead AliveCor electrocardiogram application for the screening of atrial fibrillation: a systematic review". In: *Medicine* 99.30 (2020).

[7] J. Torres-Soto and E. A. Ashley. "Multi-task deep learning for cardiac rhythm detection in wearable devices". In: *NPJ digital medicine* 3.1 (2020), p. 116.

[8] I. D. Mienye, Y. Sun, and Z. Wang. "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease". In: *Informatics in Medicine Unlocked* 18 (2020), p. 100307.

[9] D. Spathis, I. Perez-Pozuelo, S. Brage, N. J. Wareham, and C. Mascolo. “Self-supervised transfer learning of physiological representations from free-living wearable data”. In: *Proceedings of the Conference on Health, Inference, and Learning*. 2021, pp. 69–78.

[10] G. Quer, P. Gouda, M. Galarnyk, E. J. Topol, and S. R. Steinhubl. “Inter-and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults”. In: *Plos one* 15.2 (2020), e0227709.

[11] L. McInnes, J. Healy, and J. Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[12] J. H. Coote. “Recovery of heart rate following intense dynamic exercise”. In: *Experimental physiology* 95.3 (2010), pp. 431–440.

[13] P. K. Gyawali, B. M. Horacek, J. L. Sapp, and L. Wang. “Sequential factorized autoencoder for localizing the origin of ventricular activation from 12-lead electrocardiograms”. In: *IEEE Transactions on Biomedical Engineering* 67.5 (2019), pp. 1505–1516.

[14] X. Lan, D. Ng, S. Hong, and M. Feng. “Intra-inter subject self-supervised learning for multivariate cardiac signals”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 4. 2022, pp. 4532–4540.

4

HEART DISEASE DETECTION USING AN ACCELERATION-DECELERATION CURVE-BASED NEURAL NETWORK WITH CONSUMER-GRADE SMARTWATCH DATA

Cardiovascular disease (CVD) is the most important cause of morbidity and mortality worldwide. Early detection, prevention or even prediction is of pivotal importance to reduce the burden of cardiovascular disease and its associated costs. Low cost, consumer-grade smartwatches have the potential to revolutionize cardiovascular medicine by enabling continuous monitoring of heart rate and activity. When combined with machine learning(ML), the resulting large amounts of time series data hold the potential of detection, or exclusion of CVD. However, analyzing such large datasets is challenging due to the sparse presence of informative segments. Efficient selection of these segments is essential for developing predictive models for clinical deployment. The objective of this paper was to investigate the potential of an acceleration-deceleration curvebased ML model as a novel clinical indicator for the

This chapter has been published in Heliyon **10**, 21 (2024).

detection of cardiovascular disease. We used data from the ME-TIME study; 42 participants from which 21 have a cardiovascular disease and 21 are health controls. Data from each subject was normalized to decrease inter-subject variability. A neural network model aggregated predictions per week. We showed that per-subject normalization by the peak value of curves during inactivity, aggregation of model predictions over a week, and using a contrastive loss, resulted in a predictive model with 99% ± 3% specificity and 40% ± 49% sensitivity on the development set, and 100% specificity with 67% ± 47% sensitivity on the test set. Acceleration-deceleration curves are effective patterns for ruling out the presence of cardiovascular disease, but caution must be taken to properly pre-process the curves and carefully choosing a model that reduces the variability in the extracted curves.

4.1. INTRODUCTION

Cardiovascular disease (CVD), a leading cause of death [1], includes conditions like atrial fibrillation (AF) and heart failure (HF), which significantly contribute to the global morbidity- and mortality rate. To diagnose these conditions multiple tests like electrocardiograms (ECGs), ultrasound, cardiac CT and - MRI will be performed resulting in multiple hospital visits. These tests are often patient unfriendly and even potentially harmful. Additionally, significant costs are involved. In contrast, smartwatches have proven to be easy to use, comfortable to wear and have the ability to continuously measure over long periods. The primary sensor in these devices is the photoplethysmogram (PPG), which is used to estimate heart rate [2]. Heart rate variability (HRV) [3] is a strong clinical indicator of heart disease, positioning smartwatches as potential medical informative devices for consumers. This potential is further enhanced by the ongoing integration of additional sensors into smartwatches, like accelerometers, which can provide more context to heart rate (variability) data, thereby improving its predictive power.

Machine learning (ML) models are crucial to detect and predict cardiovascular outcomes from the resulting (HRV) data. These models not only aid in early detection but can also be useful in monitoring the possible progression of heart-related conditions over time and/or treatment effects. The strength of ML lies in its ability to automate the analysis of the vast amounts of HRV data. Furthermore, the data is not clinically informative to a physician at first sight, yet HRV patterns related to CVD can be effectively discovered and utilized by a ML model. Recent success in detecting atrial fibrillation (AF) using smartwatch data from raw PPG and ECG signals demonstrates the potential of wearable devices. However, these studies have primarily relied on short-term recordings [4][5][6], whereas the true potential of smartwatches lies in continuous, long-term outpatient monitoring and analysis. Similarly, heart failure detection has shown promising results [7], but these findings are also based on short ECG recordings during hospitalization.

Moreover, the raw data and algorithms used in large studies led by manufacturers [8][9][10][11] are proprietary and device-specific, making it difficult to benchmark these models in subsequent studies and impeding cross-device algorithm development for a wider range of cardiovascular disease (CVDs). In addition, some studies have utilized PPG-derived heart rate data [12][13], sometimes in combination with step counter data [14]. However, these models require ECG-based labels for training, limiting their flexibility to be adapted for other CVDs. In contrast, our method demonstrates a robust cardiovascular health indicator for both atrial fibrillation and heart failure. It only requires heart rate data—whether from PPG or ECG—and step counter data, which is readily available from commercial smartwatches. Importantly, our approach does not depend on labels from other wearables that need

to be worn or implanted simultaneously, offering greater flexibility for use in different contexts.

Entrusting the entire longitudinal data set to a machine learning model can be counterproductive if only a small proportion of the time series contains indicators for cardiovascular disease. Therefore, it is beneficial to select a subset of informative segments from the time series, guided by pre-existing cardiological knowledge. This can benefit the model's performance and provide more insights to physicians by relating the outcome of the model with patterns known from the field.

In exercise physiology an important prognosticator is the maximum aerobic capacity (VO2max) or maximum oxygen consumption. It is associated with lower cardiovascular and total mortality. Heart rate recovery (HRR) after exercise [15][16][17] is associated with VO2max and easily measurable using wearables. It refers to the speed at which the heart rate returns to a baseline level after a physical activity. A rapid HRR is typically indicative of a healthier heart and a more efficient autonomic nervous system. Conversely, a delayed HRR is associated with a higher risk of cardiovascular morbidity and mortality.

This type of controlled testing is typically not present in observational, longitudinal data as subjects are monitored in their daily free-living environment. Nonetheless, similar patterns to the HRR after exercise can occur in daily life, for example due to physical activity like exercising, climbing stairs or household chores, certain foods or beverages [18][19] and emotional responses. These occurrences essentially serve as a rudimentary gauge of cardiovascular fitness, akin to a 'poor person's exercise test', reflecting the body's response to routine stressors and activities. In addition to the HRR after exercise, heart rate acceleration capacities is associated with heart failure [20].

This paper explores patterns characterized by both an onset phase, where the heart rate increases to a peak level, and a recovery phase, where the heart rate recovers back to a baseline level. We introduce the concept of 'acceleration-deceleration (acc-dec) curves' to describe these patterns. Our contributions include a detector to extract acc-dec curves from heart rate time series data. Furthermore, we present a ML model capable of excluding cardiovascular disease (CVD) based on acc-dec curves from subjects with CVD and reference (REF) subjects without CVD, utilizing a normalization based on a step counter, weekly aggregation of curves, and a divergence-based loss for improved accuracy. We also demonstrate the clinical relevance of our model through validation on the ME-TIME [21] data set, showing its potential in enhancing early detection and monitoring of cardiovascular conditions.

4.2. MATERIALS AND METHODS

The complete ML model pipeline is illustrated in Fig. 4.1 and detailed in further sections. Briefly, acc-dec curves are extracted from the heart rate time series and normalized individually for each subject. The data is then grouped in batches where each batch contains curves from different patients and multiple weeks. The batching is required, because the outputs of the neural network are aggregated on a weekly basis in a later phase. Therefore, the number of curves must span an entire week to perform meaningful aggregation. A neural network then assigns a score between 0 and 1 to each curve in a batch, where larger than 0.5 is classified as coming from a patient with cardiovascular disease (CVD), and smaller than 0.5 is classified as healthy patient (REF). After all curves in a batch have been assigned a prediction by the network, the scores of predictions that fall in the same week, are averaged per subject together and finally a logistic layer classifies the weeks as belonging to either the REF or CVD class.

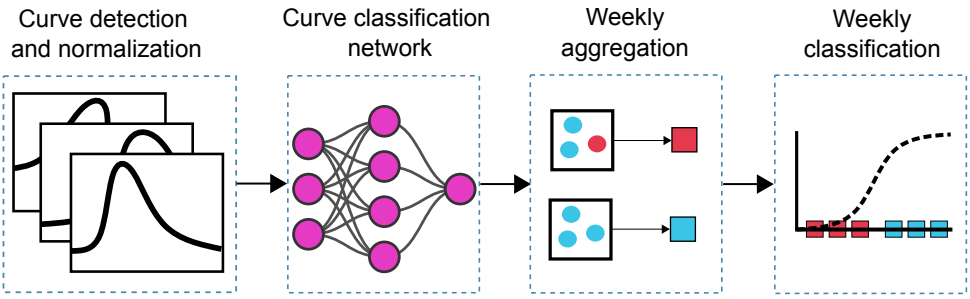


Figure 4.1.: Proposed model to detect cardiovascular disease from acc-dec curves. In the first stage, the curve detector extracts acc-dec curves and applies normalization per subject. These are then grouped in batches (both during training and deployment) and input to a neural network, which assigns a probabilistic score ranging from zero to one to each curve, illustrated as blue (REF) and red (CVD) circles. Subsequently, these predictions are grouped per subject by week, illustrated by black squares. An average of these weekly grouped predictions is then calculated, represented by blue and red squares. In the final stage a classifier is trained on the weekly aggregated predictions.

4.2.1. CURVE DETECTION AND NORMALIZATION

The data is first filtered by an eighth order Butterworth low-pass filter with critical frequency of 0.075 Hz and this filter is applied both forward and backward to cancel out phase delays/shifts introduced by the filter [22]. The critical frequency was chosen based on visual inspection compared to the raw data, ensuring that the filter preserved the overall shape of the curve while effectively attenuating higher frequency noise. Then, the detector extracts acc-dec curves from the heart rate time series by identifying three fiducial points as demonstrated in Fig. 4.2. First, the peak heart rate of the acc-dec curve is identified. This peak heart rate is the absolute heart rate value at the highest point of the acc-dec curve, with the corresponding peak point indicating its specific location. It is identified using SciPy's local maxima detection method [23], with a key parameter of this method being the prominence parameter. Prominence measures the relative height of the peak heart rate compared to the minimum heart rate in the surrounding data. It is calculated by finding the difference between the peak heart rate and the lowest heart rate value in the region between this peak and the adjacent peaks on either side.

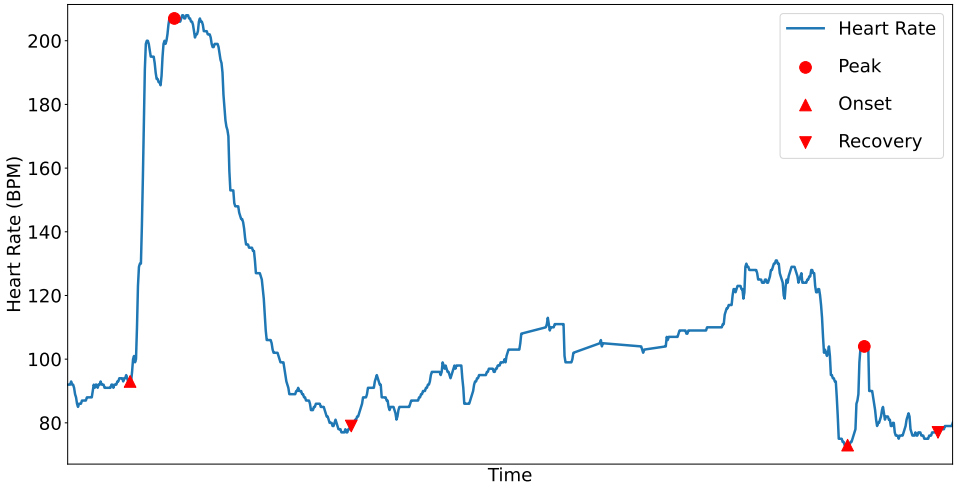


Figure 4.2.: Example of a heart rate signal in which two acc-dec curves are detected. Red circle, upward arrow and downward arrow represent peak, onset and recovery fiducial points, respectively.

Multiple acc-dec curves are identified for each subject, aligned at their peak points, and then normalized for amplitude. Two normalisations are considered: a quantile normalization, and a mean inactive-peak normalization. In the quantile normalization, for each subject, the peak

values that fall within the q-th quantile are averaged and used as normalization constant. In the mean inactive-peak normalization, for each subject, all peaks are averaged for which the subject is inactive (i.e. step counter is 0) during the time interval between the onset and recovery point of the acc-dec curve, and that is used as normalization constant.

4.2.2. CURVE CLASSIFIER

To maintain class-balance for the neural network model, during training the curves are randomly resampled, such that there is an equal number of curves originating from both REF and CVD subjects. In order to have samples of the same dimensionality, we consider acc-dec curves from 300 s before and until 600 s after the peak. Since the heart rate sample frequency is 0.2 Hz, this results in samples of dimensionality 241 ($(300+600+5)\cdot 0.2$), where the additional 5 s represents the peak itself. The acc-dec curves are then grouped in batches of 8192 curves, totalling 6 batches and used as input to a neural network.

The neural network model consists of a (acc-dec) curve classifier, which classifies each acc-dec curve as belonging to the CVD class or REF class, followed by per-subject weekly aggregation and weekly classification. During training, backpropagation starts at the weekly classification stage and flows backward through the network, reaching all the way to the curve classifier.

The curve classifier consists of two layers, where the first layer maps the 241 samples to a 20-dimensional hidden space with a tanh activation function, whereafter the final layer maps the 20-dimensional representation to a single output with a logistic activation function. This output, which varies from 0 to 1, represents a confidence score assigned by the model. A score of 1 indicates the model's highest confidence in classifying the curve as belonging to the CVD class, while a score of 0 corresponds to its association with the REF class.

4.2.3. WEEKLY AGGREGATING NEURAL NETWORK

For each subject, the outputs of the curve classifier from the same week are averaged. This average essentially serves as an indicator of the prevalence, or lack thereof, of CVD within that specific week for the individual subject. This results in a distribution representing weekly aggregated scores over all subjects. It should be emphasized that in order to enable weekly predictions, predictions based on weekly-aggregated acc-dec curves must be performed also during deployment of the model. Furthermore, it is assumed that during training all peaks from a CVD subject are considered to be CVD and all peaks from a REF subject are considered non-CVD. The same holds for the weekly aggregated scores.

Finally, a logistic regressor is trained on the weekly averaged data, again outputting a score between 0 and 1, for which, again, values between 0 and 0.5, are predicted to be non-CVD and between 0.5 and 1, predicted as CVD.

4.2.4. DIVERGENCE-BASED LOSS

The model is optimized by minimizing the cross-entropy between the weekly aggregated predictions $\hat{y}$ and the label y of the subject it originated from. Additionally, we used different regularization functions based on divergence methods that maximize the dissimilarity between predicted REF and CVD outcomes. This results in a loss function L:

$$L = L_{CE}(y, \hat{y}) + \alpha D(\hat{y}_{CVD}, \hat{y}_{REF}) \tag{4.1}$$

where L_{CE} is the cross-entropy loss, D the divergence-based regularizer and α the regularization coefficient controlling the impact of the regularizer on the overall loss function. The first regularization method aims to only separate the means and is formulated as:

$$D_{mean}(\hat{y}_{CVD}, \hat{y}_{REF}) = \frac{1}{(m(\hat{y}_{CVD}) - m(\hat{y}_{REF}))^2} \tag{4.2}$$

where $m(\hat{y})$ is the sample mean of the weekly predictions of the corresponding class.

In the second method, we separate the variances in addition to the means using a KL-divergence assuming Gaussian distributions[24]. In the univariate case this becomes:

$$L_{KL}(\hat{y}_{CVD}, \hat{y}_{REF}) = log\left(\frac{s(\hat{y}_{REF})}{s(\hat{y}_{CVD})}\right) + \frac{(s(\hat{y}_{REF}))^2 + (m(\hat{y}_{CVD}) - m(\hat{y}_{REF}))^2}{2(s(\hat{y}_{REF}))^2} - \frac{1}{2} \tag{4.3}$$

where $s(\hat{y})$ is the sample standard deviation of the weekly predictions of the corresponding class.

In contrast to the KL-divergence, we aim to maximize the divergence. Therefore in analogy with the mean divergence in eq.4.2, we take the reciprocal:

$$D_{KL} = \frac{1}{L_{KL}} \tag{4.4}$$

In the final method, we consider the contrastive loss [25]. This loss considers the Euclidean distance between all possible pairwise combinations of the weekly predictions, $\hat{y}$. Pairs belonging to the same class are identified as positive pairs, whereas those from different classes are termed negative pairs. For the positive pairs, the Euclidean distance is defined as:

$$d_k^{pos} = ||\hat{y}_i - \hat{y}_j||_2^2 \tag{4.5}$$

where d_k^{pos} contains the Euclidean distances of all positive pairs and similarly d_l^{neg}, denotes those of all negative pairs. The objective is to keep positive pairs closely together, while simultaneously separate negative pairs, thereby achieving a clear distinction between the classes. The contrastive loss is given as:

$$D_{contrastive} = \frac{1}{N^{pos}} \sum_{k=1}^{N^{pos}} \left(d_k^{pos}\right)^2 + \frac{1}{N^{neg}} \sum_{l=1}^{N^{neg}} \max\left(0, \beta - \left(d_l^{neg}\right)^2\right). \quad (4.6)$$

The first term of the equation aims to reduce the Euclidean distances between positive pairs, denoted by N^{pos} (i.e. the total number of these pairs). Conversely, the second term seeks to increase the distance between negative pairs, where N^{neg} represents the total number of negative pairs. Additionally, the introduction of a margin hyperparameter, β, combined with the use of the max operator, plays a crucial role. It limits the penalty to distances up to β, allowing the model to concentrate on "hard examples" (i.e. negative and positive pairs that are much alike). This focus is beneficial as it prevents unnecessary optimization of samples that are already satisfactorily positioned, where further re-positioning wouldn't enhance the model's performance.

4.2.5. ME-TIME DATA SET

The ME-TIME study (registered at clinicaltrials.gov with ID NCT05802563) is an observational, longitudinal study conducted at the Haga teaching hospital and is approved by the Institutional Review Board (or Ethics Committee) of METC-LDD (protocol code NL73708.058.20) for studies involving humans. The study, which included a 12-week observation period, aims to develop methods and algorithms for remote detection and prevention of cardiovascular disease using Fitbit smartwatches[21]. The observation period was 12 weeks. Tables 4.1 and 4.2 show the characteristics of the development set, used to train and tune the neural network model, as well as the test set.

The REF group consists of 15 subjects in the train and 6 subjects in the test set. The CVD group is divided into train and test in a similar way.
In comparison to the REF group, the CVD group consists of slightly older individuals with higher BMI and a higher prevalence of hypertension. The CVD group consists of subjects with some type of atrial fibrillation (AF) or heart failure (HF).

Table 4.1.: Characteristics of the development set used to train and tune a neural network model.

Characteristic		REF	HF	PersAF	PermAF	HF+AF
Participants, n		15	8	3	4	15
Age, y	18-39	6	0	0	0	6
	40-54	5	1	1	0	2
	55-64	4	3	0	0	3
	>65	0	4	2	4	10
Sex	Male	4	6	3	2	11
	Female	11	2	0	2	4
BMI	18.5–24.9	7	2	0	1	3
	25–29.9	5	2	2	2	6
	>30	3	4	1	1	6
Diabetes	Yes	0	4	0	2	6
	No	15	4	3	2	10
Smoking	Yes	0	5	0	3	8
	No	15	3	3	1	7
Hypertension	Yes	1	5	1	3	9
	No	14	3	2	1	6
Device	Charge 5	10	5	2	3	10
	Inspire 2	5	3	1	1	5

Table 4.2.: Characteristics of the test set used to evaluate the performance of the trained and tuned model.

Characteristic		REF	HF	PersAF	PermAF	HF+AF
Participants, n		6	4	1	1	6
Age, y	18-39	4	0	0	0	0
	40-54	1	1	0	0	1
	55-64	1	1	0	0	1
	>65	0	2	1	1	4
Sex	Male	3	3	1	0	4
	Female	3	1	0	1	2
BMI	18.5–24.9	3	2	0	0	2
	25–29.9	3	1	1	1	3
	>30	0	1	0	0	1
Diabetes	Yes	0	1	1	0	2
	No	6	3	0	1	4
Smoking	Yes	0	1	0	0	1
	No	6	3	1	1	5
Hypertension	Yes	0	2	1	0	3
	No	6	2	0	1	3
Device	Charge 5	5	4	0	1	5
	Inspire 2	1	0	1	0	1

4.2.6. EVALUATION

The measures reported are Sensitivity (also known as True Positive Rate) and Specificity (also known as True Negative Rate), which are calculated on a per-subject basis and defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.8}$$

where where True Positive (TP) represents the number of weeks correctly classified as CVD. False Negative (FN) the number of weeks misclassified as REF. False Positive (FP) the number of weeks misclassified as CVD. True Negative (TN) the number of weeks correctly classified as REF. This results in a sensitivity and specificity for each subject after which the population average and standard deviation across subjects is reported.

4.2.7. MODEL TRAINING

Parameters such as the margin and regularization coefficient mentioned in Section 4.2.4) cannot be determined through the neural network's learning process and are called hyperparameters. Instead, these are found through trial-and-error by trying out different values. The best hyperparameter values are determined by selecting the configuration with the highest average accuracy across folds in a leave-6-subjects-out stratified cross-validation approach on the training set only. This procedure is further elaborated in Appendix B.

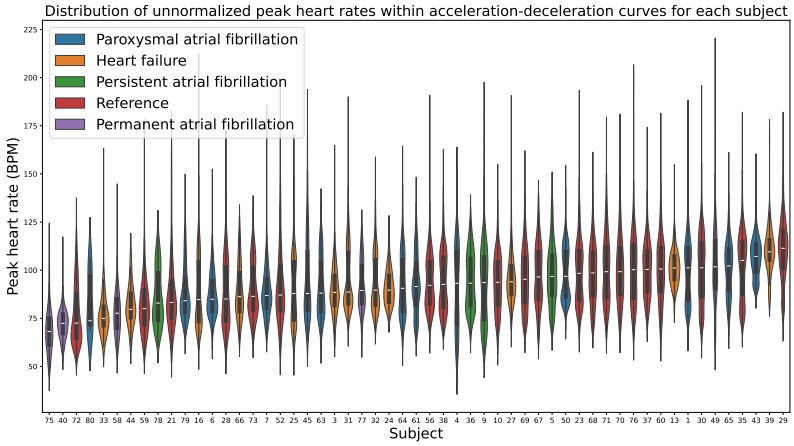
4.3. RESULTS

4.3.1. INTER-SUBJECT AND INTRA-SUBJECT VARIABILITY

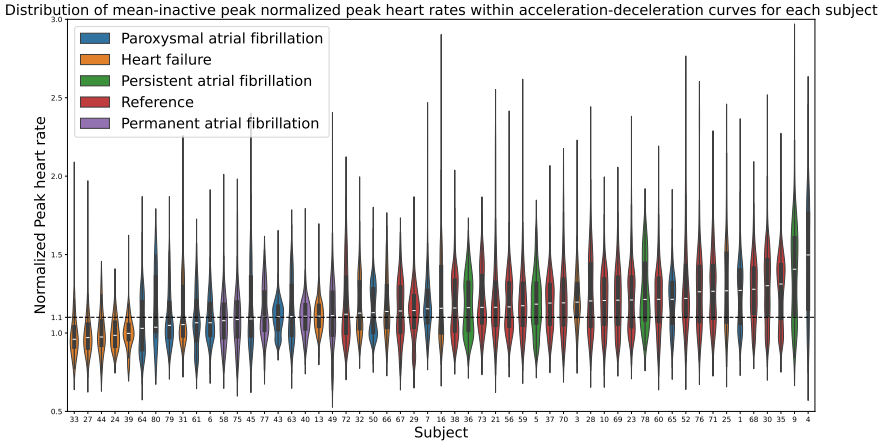
The results illustrated in Fig. 4.3 highlight significant variations in the distribution of peak heart rates within the acc-dec curves among participants. For this analysis, we also include a paroxysmal AF group. These variations are evident not only between different subjects (inter-subject variability) but also within the measurements from individual subjects (intra-subject variability). By applying the mean inactive peak normalization and sorting the subjects by median peak heart rate, it becomes apparent in Fig. 4.3 that the normalized peak heart rate serves as an informative feature for CVDs. Specifically, values approximately above 1.1 (10 percent above the baseline peak heart rate) generally correspond to reference subjects. Furthermore, values below the 1.0 baseline seem to correspond to heart failure patients. These findings underscore the complex nature of cardiovascular responses as captured by acc-dec curves.

We investigated the impact of the normalisation on the difference in acc-dec curves between the CVD and REF group. Fig. 4.4 shows average acc-dec curves, with and without normalization. Here, the distance between the REF and CVD curves appears to be larger when using the normalization by the mean inactive-peak. As we strive towards differentiating between these groups, we selected for this normalization scheme. Appendix B.2 demonstrates that other quantile values perform worse. We also investigated the impact of the prominence and activity level in Appendix B.3.

This revealed that, using acc-dec curves with mean inactive-peak normalization, a prominence of 20 BPM and activity levels above 20 steps, the distance between the average REF and CVD curve appears to be largest, among the methods that we investigated. Therefore we use this configuration to train a ML model.



(a)



(b)

Figure 4.3.: Unnormalized (top, a) and mean-inactive peak normalized (bottom, b) distribution of peak heart rates per subject. A: Distribution of unnormalized peak heart rate in acc-dec curves for each subject. B:Distribution of mean-inactive peak normalized peak heart rate across acceleration-deceleration curves for each subject.

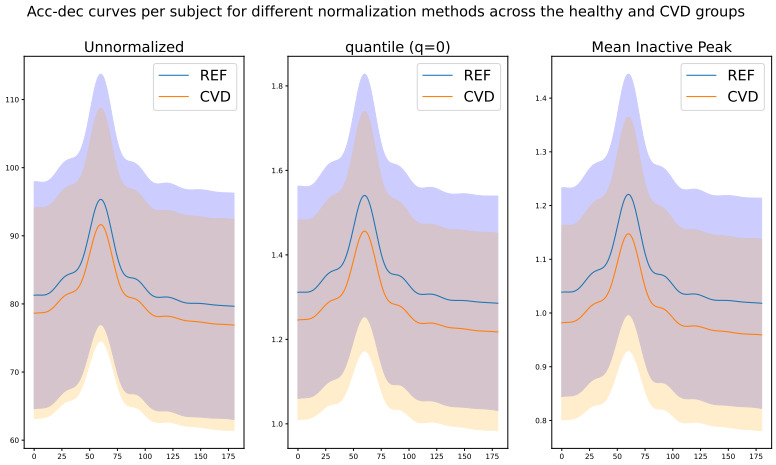


Figure 4.4.: Unnormalized (left), quantile normalization (middle, using q = 0, effectively using the smallest peak) and mean inactive-peak normalization (right).

4.3.2. ABLATION STUDY OF PREDICTIVE MODEL

Using these hyperparameters, we developed ML models to classify acc-dec curves of participants into CVD patients or REF individuals. We performed an ablation study (shown in Table 4.3) starting with a naive baseline, which only consists of the curve-based classifier without normalization or weekly aggregation/classification from Figure 4.1. Then we have sequentially added the mean inactive-peak normalization and aggregation, and the last component added is the divergence-based regularization, for which we investigate three alternatives: mean divergence, KL divergence and contrastive loss. By doing so, the impact of each of these additions on the model's performance is quantified.

The results in Table 4.3 show that the naive model performs poorly.

Applying mean inactive-peak improves the specificity of the model. This enhancement becomes more pronounced when the aggregation technique are also implemented.

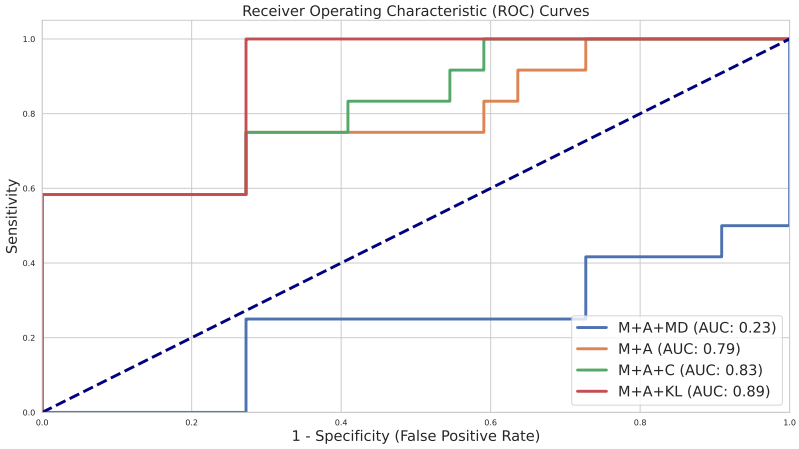
On top of normalisation and aggregation, the addition of the divergence losses is considered, resulting in three alternatives. When mean divergence is applied, the model's performance significantly deteriorates. Closer examination of the mean divergence model's predictions in Fig. 4.5b reveals that it always classifies all samples as CVD, resulting in 100% sensitivity at 0% specificity.

The specificity of the model benefits most from contrastive regularization. While each added component increases specificity, it simultaneously leads to a reduction in sensitivity. Considering the high

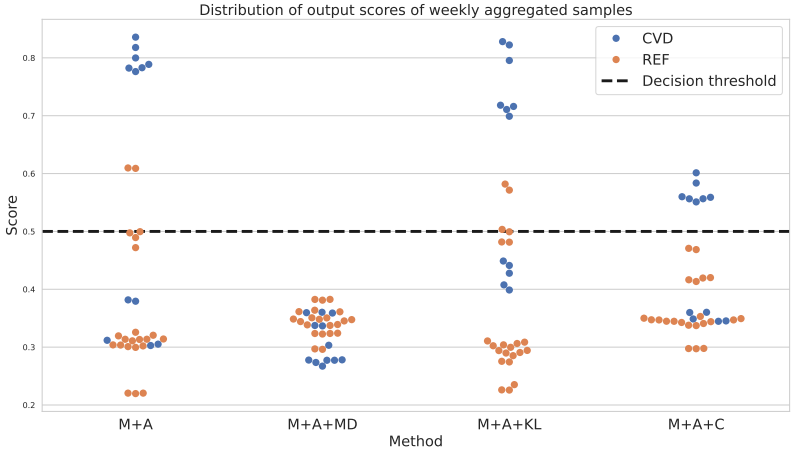
Table 4.3.: Ablation analysis of average model performance across subjects with incremental addition of: mean inactive-peak (M) normalization, Weekly Aggregation (A), Mean Divergence (MD), KL Divergence (KL), Contrastive Loss (CL). Naive model consists of curve-based classifier only. Standard deviations are provided in parentheses to illustrate the variability across subjects.

Set	Metric	Naive	M	M+A	M+A+MD	M+A+KL	M+A+CL
Development	Sensitivity	44 (28)	60 (25)	44 (48)	20 (40)	51 (48)	40 (49)
	Specificity	63 (18)	61 (19)	88 (16)	80 (40)	89 (27)	99 (3)
Test	Sensitivity	61 (18)	64 (19)	67 (48)	0 (0)	67 (47)	67 (47)
	Specificity	53 (19)	62 (12)	83 (19)	100 (0)	79 (37)	100 (0)

specificity, the model is very good at detecting reference cases when they are present. However, the low average for the sensitivity (43 %), and its high standard deviation (48 %), indicate that the model is not consistently good at identifying CVD cases as seen in Figure 4.5b. Specifically, for some CVD test subjects, it classifies almost all positive weeks correctly, while in others, it fails to classify any weeks correctly. Thus, the model's positive predictive performance closely resembles that of random guessing, underscoring its lack of consistency and reliability in detecting CVD cases. In addition, contrastive regularization results in an accuracy of 85%, positive predictive value of 100% and negative predictive value of 82%. Finally, KL divergence performs better than mean divergence, however it does not improve upon using contrastive regularization or even only normalisation and aggregation. However, Figure 4.5a reveals that it performs best in terms of the area under the receiver operator curve.



(a)



(b)

Figure 4.5.: ROC curves for aggregation methods (top, a) and output scores of weekly aggregated predictions on the test set (bottom, b). Colors represent the true labels: blue for CVD and orange for REF. The dashed line indicates the decision threshold of 0.5, with scores below 0.5 classified as REF and scores above 0.5 classified as CVD.

4.4. DISCUSSION & CONCLUSION

We have introduced acc-dec curves in combination with a ML model to demonstrate high specificity, reliably identifying the absence of cardiovascular disease in longitudinal data acquired from consumer-grade wearables. Deployed as a "rule out" modality this can be very powerful in clinical practise. Currently, many patients visiting a physician with complaints require further testing to diagnose the disease causing these complaints. Medical tests are often designed to have a high specificity to optimize the negative predictive value for the disease. For example, cardiac CT has a 99 percent specificity for the presence of coronary calcification [26] making it a powerful tool to rule out coronary artery disease and strongly modifying cardiovascular risk. However, performing an expensive scan (which also uses harmful ionizing radiation and contrast causing patients' discomfort) only provides information on the current status (a 'snap shot'). In contrast, a monitoring system based on this ML model can be used by the patients for an extensive period of time, has no harmful effects on the patient and it also has the potential to avoid unnecessary visits to a cardiologist for patients who have symptoms that are initially suspected to be related to heart problems, but after assessment are determined not to be. Our method on predicting CVD also has a high specificity, for which a similar clinical efficiency reasoning can be followed.

Secondly it is shown that it is important that the acc-dec curves are normalized per-subject based on scaling by the step count. In our analysis, scaling the curves using the average peak value observed during inactivity proves to be effective. Our study involved comparing acc-dec curves between healthy individuals and CVD patients, focusing on three distinct ranges based on step counter readings: no steps (0), low activity (1–20 steps), and higher activity (exceeding 20 steps). Curves recorded, during periods of higher activity appeared to be more effective in differentiating between the two groups.

Thirdly, we show that it is beneficial to aggregate model predictions of curves that fall within the same week. The effectiveness may be attributed to its ability to mitigate the large inter- and intra-subject variability in acc-dec curves. Such variabilities could challenge machine learning models to generalize effectively across a diverse population and must generally be accounted for. Other time intervals could also be explored and in general, incorporating information about time (of the day) could be beneficial.

Fourthly, we show that it is effective to use a divergence-based loss that ensures that the model finds a representation where data from the two classes, healthy and cardiovascular disease, are far apart. Since originally the mean divergence and KL-divergence are suited for minimization problems, we took the reciprocal to maximize the distance between the two classes. However, other functions that transform

divergence-based losses into maximization methods could also be explored such as negation. Compared to the reciprocal functions, the contrastive loss performs best. This difference may be because the mean and KL divergence only optimize for summary statistics (mean and variance), whereas contrastive loss penalizes pairwise predictions, thus operating at a more granular level. Other contrastive functions such as the triplet loss could also be explored.

The results in Table 4.3 and Fig. 4.5b show that the models are useful to rule out the presence of CVD, due to the high specificity (100%), but cannot detect CVD very well due to the low sensitivity (65% ± 47%). Because the models have this characteristic, they can be a very strong tool in medicine, and especially in cardiology, as clinicians usually work "per exclusionem" or by ruling out a diagnosis. Therefore may be of significant clinical importance.

Our study is limited by the fact that the sample size on which the model is tested is rather small.

Lastly, we have considered the classification performance of the weekly aggregated data. By utilizing subject labels to give predictions on a per-subject basis (i.e. aggregating over multiple weeks), this approach might allow for more consistently accurate predictions.

Future research should be directed towards external validation of the model on various datasets (e.g. multi-center study) from different smartwatch devices to confirm the data. The goal should be to safely implement these models in clinical practice and to improve medical care. Furthermore, more diagnoses should be trained using the methods we propose and this work should be directed towards diseases with a high prevalence and incidence such as coronary artery disease.

Furthermore, successful integration of such models into clinical practice comes with challenges. For example, seamless integration with existing healthcare systems is essential for the model to provide real-time, actionable insights. Furthermore, healthcare providers would need training on interpreting model outputs and effectively integrating them into patient care workflows. Patient compliance also plays a critical role, in order to capture sufficient acc-dec curves. Additionally, a proper database infrastructure is required to store and manage the data effectively, and cloud platforms are a suitable option for this purpose. However, the use of cloud platforms introduces privacy considerations that must be addressed to ensure the protection of sensitive patient information. Addressing these challenges is vital to ensure smooth adoption and widespread implementation of the proposed model.

Future research should be directed towards external validation of the model on various datasets (e.g. multi-center study) from different smartwatch devices to confirm the data. The goal should be to safely implement these models in clinical practice and to improve medical care. Furthermore, more diagnoses should be trained using the methods

we propose and this work should be directed towards diseases with a high prevalence and incidence such as coronary artery disease.
In summary, this study demonstrates the effectiveness of acc-dec curves acquired from a consumer-grade smartwatch combined with machine learning, offering a non-invasive, efficient, and powerful tool to rule out cardiovascular disease. This shows promise of transforming clinical practice and enhancing patient care through advanced, ML-driven methodologies, contributing to a new era of remote patient monitoring.

REFERENCES

[1] A. Shi, Z. Tao, P. Wei, and J. Zhao. “Epidemiological aspects of heart diseases”. In: *Experimental and therapeutic medicine* 12.3 (2016), pp. 1645–1650. issn: 1792-0981.

[2] J. Torres-Soto and E. A. Ashley. “Multi-task deep learning for cardiac rhythm detection in wearable devices”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–8. issn: 2398-6352.

[3] J. F. Thayer, S. S. Yamamoto, and J. F. Brosschot. “The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors”. In: *International journal of cardiology* 141.2 (2010), pp. 122–131.

[4] J. Bacevicius, Z. Abramikas, E. Dvinelis, D. Audzijoniene, M. Petrylaite, J. Marinskiene, J. Staigyte, A. Karuzas, V. Juknevicius, R. Jakaite, *et al.* “High specificity wearable device with photoplethysmography and six-lead electrocardiography for atrial fibrillation detection challenged by frequent premature contractions: doubleCheck-AF”. In: *Frontiers in cardiovascular medicine* 9 (2022), p. 869730.

[5] J. Ramesh, Z. Solatidehkordi, R. Aburukba, and A. Sagahyroon. “Atrial fibrillation classification with smart wearables using short-term heart rate variability and deep convolutional neural networks”. In: *Sensors* 21.21 (2021), p. 7233.

[6] L. Zhu, V. Nathan, J. Kuang, J. Kim, R. Avram, J. Olgin, and J. Gao. “Atrial fibrillation detection and atrial fibrillation burden estimation via wearables”. In: *IEEE Journal of Biomedical and Health Informatics* 26.5 (2021), pp. 2063–2074.

[7] J.-m. Kwon, Y.-Y. Jo, S. Y. Lee, S. Kang, S.-Y. Lim, M. S. Lee, and K.-H. Kim. “Artificial intelligence-enhanced smartwatch ECG for heart failure-reduced ejection fraction detection by generating 12-lead ECG”. In: *Diagnostics* 12.3 (2022), p. 654.

[8] M. V. Perez, K. W. Mahaffey, H. Hedlin, J. S. Rumsfeld, A. Garcia, T. Ferris, V. Balasubramanian, A. M. Russo, A. Rajmane, L. Cheung, *et al.* “Large-scale assessment of a smartwatch to identify atrial fibrillation”. In: *New England Journal of Medicine* 381.20 (2019), pp. 1909–1917.

[9] S. A. Lubitz, A. Faranesh, C. Selvaggi, S. Atlas, D. D. McManus, D. E. Singer, S. Pagoto, A. Pantelopoulos, and A. Foulkes. “Detection of atrial fibrillation in a large population using wearable devices: the Fitbit Heart Study”. In: *Circulation*. Vol. 144. 25. LIPPINCOTT WILLIAMS & WILKINS TWO COMMERCE SQ, 2001 MARKET ST, PHILADELPHIA . . . 2021, E570–E571.

[10] P. Badertscher, M. Lischer, D. Mannhart, S. Knecht, C. Isenegger, J. D. F. de Lavallaz, B. Schaer, S. Osswald, M. Kühne, and C. Sticherling. “Clinical validation of a novel smartwatch for automated detection of atrial fibrillation”. In: *Heart Rhythm O2* 3.2 (2022), pp. 208–210.

[11] Y. Guo, H. Wang, H. Zhang, T. Liu, L. Li, L. Liu, M. Chen, Y. Chen, and G. Y. Lip. “Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: a report from the huawei heart study”. In: *JACC: Asia* 1.3 (2021), pp. 399–408.

[12] D. Hiraoka, T. Inui, E. Kawakami, M. Oya, A. Tsuji, K. Honma, Y. Kawasaki, Y. Ozawa, Y. Shiko, H. Ueda, *et al.* “Diagnosis of atrial fibrillation using machine learning with wearable devices after cardiac surgery: algorithm development study”. In: *JMIR Formative Research* 6.8 (2022), e35396.

[13] J. Wasserlauf, C. You, R. Patel, A. Valys, D. Albert, and R. Passman. “Smartwatch performance for the detection and quantification of atrial fibrillation”. In: *Circulation: Arrhythmia and Electrophysiology* 12.6 (2019), e006834.

[14] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone, *et al.* “Passive detection of atrial fibrillation using a commercially available smartwatch”. In: *JAMA cardiology* 3.5 (2018), pp. 409–416.

[15] C. R. Cole, E. H. Blackstone, F. J. Pashkow, C. E. Snader, and M. S. Lauer. “Heart-rate recovery immediately after exercise as a predictor of mortality”. In: *New England journal of medicine* 341.18 (1999), pp. 1351–1357.

[16] S. Nanas, M. Anastasiou-Nana, S. Dimopoulos, D. Sakellariou, G. Alexopoulos, S. Kapsimalakou, P. Papazoglou, E. Tsolakis, O. Papazachou, C. Roussos, *et al.* “Early heart rate recovery after exercise predicts mortality in patients with chronic heart failure”. In: *International journal of cardiology* 110.3 (2006), pp. 393–400.

[17] S. I. Nissinen, T. H. Mäkikallio, T. Seppänen, J. M. Tapanainen, M. Salo, M. P. Tulppo, and H. V. Huikuri. “Heart rate recovery after exercise as a predictor of mortality among survivors of acute myocardial infarction”. In: *American Journal of Cardiology* 91.6 (2003), pp. 711–714.

[18] A. A. M. Duarte, C. Mostarda, M. C. Irigoyen, and K. Rigatto. “A single dose of dark chocolate increases parasympathetic modulation and heart rate variability in healthy subjects”. In: *Revista de Nutrição* 29 (2016), pp. 765–773.

[19] K. A. Sauder, E. R. Johnston, A. C. Skulas-Ray, T. S. Campbell, and S. G. West. “Effect of meal content on heart rate variability and cardiovascular reactivity to mental stress”. In: *Psychophysiology* 49.4 (2012), pp. 470–477.

[20] W. Hu, X. Jin, P. Zhang, Q. Yu, G. Yin, Y. Lu, H. Xiao, Y. Chen, and D. Zhang. “Deceleration and acceleration capacities of heart rate associated with heart failure with high discriminating performance”. In: *Scientific reports* 6.1 (2016), p. 23617.

[21] A. Naseri, D. Tax, P. van der Harst, M. Reinders, and I. van der Bilt. “Data-efficient machine learning methods in the ME-TIME study: Rationale and design of a longitudinal study to detect atrial fibrillation and heart failure from wearables”. In: *Cardiovascular Digital Health Journal* 4.6 (2023), pp. 165–172.

[22] E. Mejía-Mejía, J. M. May, and P. A. Kyriacou. “Effect of filtering of photoplethysmography signals in pulse rate variability analysis”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 5500–5503.

[23] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. doi: 10.1038/s41592-019-0686-2.

[24] L. Pardo. *Statistical inference based on divergence measures*. CRC press, 2018.

[25] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. “A tutorial on energy-based learning”. In: *Predicting structured data* 1.0 (2006).

[26] D. Andreini, G. Pontone, A. L. Bartorelli, P. Agostoni, S. Mushtaq, E. Bertella, D. Trabattoni, G. Cattadori, S. Cortinovis, A. Annoni, *et al.* “Sixty-four–slice multidetector computed tomography: an accurate imaging modality for the evaluation of coronary arteries in dilated cardiomyopathy of unknown etiology”. In: *Circulation: Cardiovascular Imaging* 2.3 (2009), pp. 199–205.

5

IMPROVING PERFORMANCE OF HEART RATE TIME SERIES CLASSIFICATION BY GROUPING SUBJECT

Unlike the more commonly analyzed ECG or PPG data for activity classification, heart rate time series data is less detailed, often noisier and can contain missing data points. Using the BigIdeasLab_STEP dataset, which includes heart rate time series annotated with specific tasks performed by individuals, we sought to determine if general classification was achievable.

Our analyses showed that the accuracy is sensitive to the choice of window/stride size. Moreover, we found variable classification performances between subjects due to differences in the physical structure of their hearts. Various techniques were used to minimize this variability. First of all, normalization proved to be a crucial step and significantly improved the performance. Secondly, grouping subjects and performing classification inside a group helped to improve performance and decrease inter-subject variability. Finally, we show that including handcrafted features as input to a deep learning (DL) network improves the classification performance further.

Together, these findings indicate that heart rate time series can be utilized for classification tasks like predicting activity. However, normalization or grouping techniques need to be chosen carefully to minimize the issue of subject variability.

This chapter has been accepted and presented at BNAIC (2023).

5.1. INTRODUCTION

In recent years, wearable devices and smartwatches have been equipped with more sensors, including electrocardiogram (ECG) and photoplethysmography (PPG) sensors, for the estimation of heart rate and heart rhythm [1]. These developments enable us to collect long-term heart rate time series data of a subject's heart rate in beats per minute (BPM). In the research community, there are many papers that attempt to perform classification using ECG or PPG data.

While ECG and PPG data shows each heartbeat's characteristics in detail, heart rate data summarizes this based on the time elapsed between heartbeats. For heart rate, we receive a single measurement, representing beats per minute, at regular intervals—like once every few seconds. This is therefore a more challenging signal to perform classification tasks on. Research on the analysis and usage of heart rate time series has been performed for example for cardiovascular risk detection[2][3]) and sleep analysis [4].

In this paper, we will look into the classification of heart rate time series data to predict different activities a subject is doing. This is interesting to test because it would imply that we can use heart rate data in the future for more complex classification problems, like heart disease detection. We will make use of the BigIdeasLab_STEP[5] dataset which contains annotated heart rate time series of subjects performing different activities.

Magure et al. perform activity classification[6] which was mostly possible due to the fact that the accelerometer was placed at strategic places to identify specific movements and the subjects all had the same age and fitness level. In contrast, Bent and Dunn[5] conducted a study involving subjects of varying skin tones performing different physical activities while wearing multiple smartwatches. Their findings revealed no statistically significant difference in accuracy across skin tones. However, there were notable increases in error during physical activity compared to rest. Specifically, the absolute error during physical activity was on average 30% greater than during periods of rest.

However, constructing a unified classification model is challenging due to the diverse characteristics between devices and subjects. Therefore we propose to group similar subjects together and construct a model for each group.

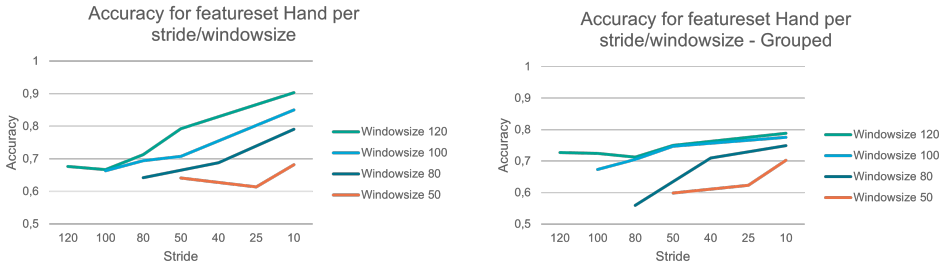
5.2. RESULTS

The BigIdeasLab_STEP dataset contains around 13 minutes of heart rate time series data per subject. This dataset is annotated with the activity a subject is performing. The activities were: resting, breathing, performing an activity (walking), resting after the activity and typing. The data is split up into windows of fixed size and a specific stride is

used between each window. A window size refers to the number of consecutive samples one takes from a certain start point. The stride indicates the number of samples the start point is shifted for the next window. A more detailed description of the dataset can be found in the methods and section C.1.

5.2.1. COMPARISON OF DIFFERENT WINDOW AND STRIDE SIZES

First, we investigate the influence of varying window and stride sizes. For that, we trained a Support Vector Machine (SVM) on windows sizes of 50, 80, 100 and 120 and stride sizes of 10, 25, 40, 50, 80, 100 and 120. A short explanation of the SVM can be found in the methods and appendix C.2. The input data used was the raw time series. We performed the experiment twice. During the first time, we used a train and test set where some windows of a person were in the train set and some were in the test set. The second time we only used a train and test set where all the windows of a person were either in the train or the test set (resulting in a "leave-subject-out" validation procedure). The results for the first two experiments can be found in Figure 5.1.



(a) Random split

(b) Leave-subject-out

Figure 5.1.: Accuracy when training an SVM using (a): a random split (windows of the same subject, both in the train and test set) and (b): when windows of subjects are either in the training or in the test set (leave-subject-out validation procedure). This is inspected for different window and stride sizes. For random splitting, the accuracy increases as the window size increase and stride size decrease, whereas for the leave-subject-out procedure, the accuracy seems to converge to one point. The achieved accuracies are plotted on the y-axis and the stride sizes are on the x-axis. The different window sizes are represented by different coloured lines. 120 (green), 100 (light blue), 80 (dark blue) and 50 (orange).

Both figures clearly illustrate that with every color-coded line,

representing different window sizes, the accuracy increases as the stride size decreases. As the stride size decreases, there is an increase in the sample size. However, even though these windows become more dependent (due to larger overlap) with the reduction in stride size, the effectively larger sample size still enhances performance. Moreover, as the window sizes get larger, the accuracy also gets higher. However, there is a difference between the two figures. For the "leave-subject-out" validation procedure, the accuracies seem to converge to one point or at least stabilise, whereas in the random split scenario, the lines show an overall increasing trend.

5.2.2. THE EFFECT OF CLUSTERING SUBJECTS

Although the classification performance shown in the previous section is reasonable, it is known that these data show large inter-person variability due to physical differences between subjects [7]. To see if more personalised models improve performance, we cluster the subjects based on various metrics. The first metric: the average heart rate (in BPM) of every activity resulted in five values per person. These five values represented a time series of five points in the order as the activities performed: rest, breath, activity, rest, and type. When we clustered these time series per person with different resulting numbers of clusters, a cluster assignment as in Figure 5.2 was achieved.

To determine whether there were differences between the cluster groups, we trained an SVM on one cluster while another cluster was used as a testing set. The combinations and the corresponding scores achieved are represented in Table 5.1.

Train x / Test y	Averaged balanced accuracy
Train 1 / Test 2	.73
Train 1 / Test 5	.40
Train 1 / Test 6	.57
Train 5 / Test 6	.36
Train 5 / Test 3	.44

Table 5.1.: Accuracies of training an SVM and using subjects of one cluster as training set and subjects of another cluster as testing set. The numbers indicate the clusters in Figure 5.2 counted from left to right. Similar clusters achieve higher accuracy than more dissimilar ones.

We can observe in this table that the clusters that look similar (eg. cluster 1 and 2) achieve a cross-cluster better performance than clusters that look more dissimilar (eg. 1 and 5). This suggests that there exists inter-subject variability in this dataset.
To investigate the existence of variability within a cluster, we trained an

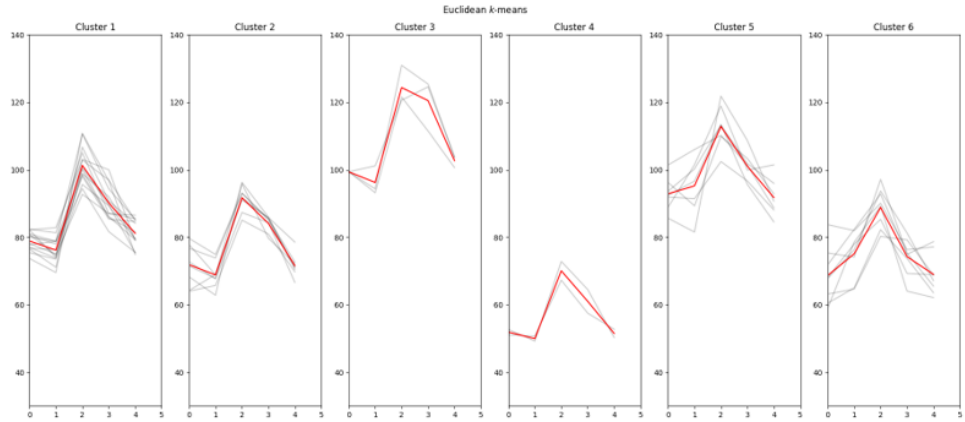
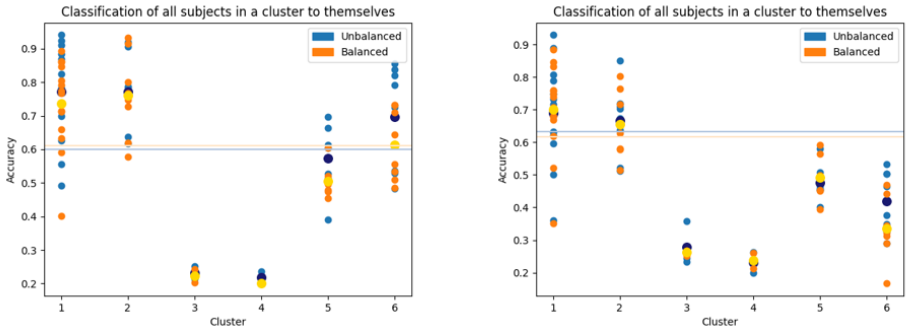


Figure 5.2.: A cluster assignment with the number of clusters equal to 6 using a time series of a subject's mean BPM per activity using the TimeSeriesKmeans clustering procedure[8]. Subplots from left to right represent the six different clusters and the subjects included. Grey lines represent the individual time series and thus represent a single subject. Red lines are the averages of the time series in the cluster. The x-axis shows the different activities numbered from 0 to 4 and the y-axis shows the heart rate in BPM.

SVM on all the data in a cluster except for one subject, which was used for testing. We performed this for every cluster and for every combination inside a cluster. We considered two different standardization methods namely 'Feature' and 'Data' standardization. In Feature standardization, z-score standardization is applied on the features after windowing and feature generation. In Data standardization, z-score standardization is applied on the original heart rate time per person, whereafter windowing and feature generation is performed. The mean and standard deviation used for standardising the training data are also used for the standardization of the features in the testing data. The results of both methods can be found in Figures 5.3a and 5.3b.

First, these figures show us that the Feature standardization case is performing better. Next, we see that in three out of four (larger) clusters, the average accuracy within a cluster is higher than the SVM when no clustering of subjects is done. Note that clusters 3 and 4 contain an insufficient number of samples to provide an accurate representation.

Next, we conducted an additional experiment to investigate if the clustering could be improved by using multiple features instead of only the mean heart rate per activity. To test this, we evaluated the within-cluster accuracies using different methods of clustering. The two



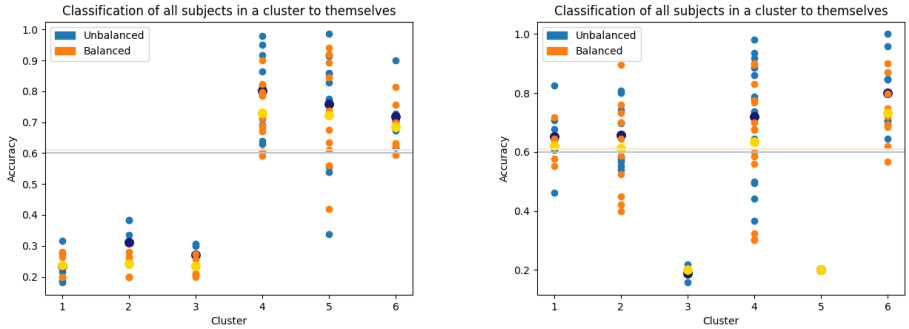
(a) Feature standardization (b) Data standardization

Figure 5.3.: Results of accuracies within a cluster for the Feature and Data standardization methods when training an SVM with the leave-one-subject out validation procedure. Yellow and dark blue points represent the mean per cluster and horizontal lines represent the performance of the SVM when no clustering is performed. For yellow/orange points, balanced accuracy was used and for light/dark blue, unbalanced/normal accuracy. With Feature standardization (a), three of the four larger clusters have a mean accuracy higher than an SVM trained on all the data. With Data standardization (b), only two of the four larger clusters, have a mean accuracy higher than an SVM trained on all the data.

different methods we investigated were the use of temporal features and the use of statistical features instead of mean heart rate. Statistical and temporal features are the features generated by TSFEL[9]. The results can be seen in Figures 5.4a and 5.4b.

These figures show that the statistical features are better for clustering than the temporal features. In all large clusters, it achieves better performance than the SVM trained when no clustering is performed. In the temporal case, this is only 3 out of 4 just like with the mean BPM clustering method.

To demonstrate that it can also help with previously unseen samples, we conducted several additional experiments using the leave-subject-out procedure. We used the training set for generating the clustering model and cluster assignment, as well as to train a model for each cluster. The test set was used in two different ways. The first approach was per-window classification. With this approach, a window of a test subject was obtained, the corresponding cluster was determined, and the model associated with that cluster was used to classify the window. The results of this approach can be seen in the first column of Table 5.2.



(a) Temporal features (b) Statistical features

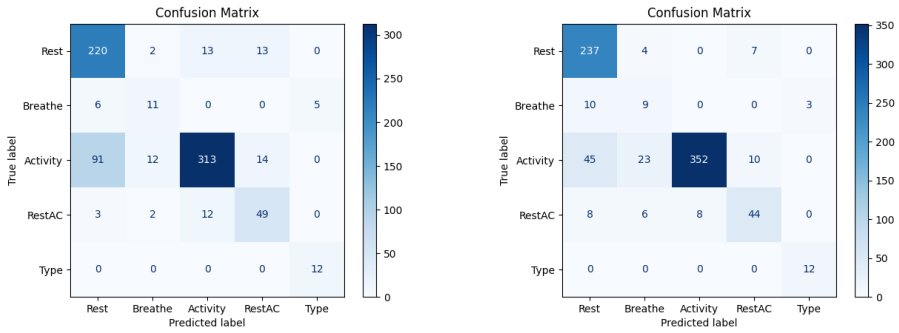
Figure 5.4.: Results of accuracies within a cluster for the Feature standardization method using temporal and statistical features and training an SVM with the leave-one-subject out validation procedure. Yellow and dark blue points represent the mean per cluster and horizontal lines represent the performance of the SVM when no clustering is performed. For yellow/orange points, balanced accuracy was used and for light/dark blue, unbalanced/normal accuracy. With temporal features (a), three of the four larger clusters have a mean accuracy higher than an SVM trained on all the data. With statistical features (b), all four clusters have a mean accuracy higher than an SVM trained on all the data.

	Per-window	Per-subject
6 clusters	.46	.74
5 clusters	.50	.63
4 clusters	.56	.75
3 clusters	.68	.72

Table 5.2.: Achieved classification accuracies when using different numbers of resulting clusters and clustering techniques to find a cluster model for activity prediction. Grouping the subjects in 4 clusters and using the per-subject method achieves the highest accuracy.

The second approach was to apply the same personalised classifier to all windows of one test subject, the per-subject approach. To do so, we used the clustering model to determine to which cluster each single window of a test subject belongs to. After this, the cluster with the

highest number of assigned windows was used to obtain the model for classifying all windows of a specific subject. The result of this experiment is presented in the second column of Table 5.2. The per-subject method achieves higher accuracies than the per-window classification method. It achieves an accuracy of 0.71 while the per-subject method with 4 clusters achieves 0.75. In the next paragraph, we delve more into the differences in prediction between both methods, rather than solely examining the achieved accuracies.
The confusion matrices of the two methods can be found in Figure 5.5. We can see a prominent difference within misclassifications occuring between the two largest classes (Rest and Activity). The per-subject approach exhibits much less misclassifications between these two classes compared to the per-window method. Depending on the application, misclassifications between these very different classes is more severe than misclassifications between similar classes (Rest vs Breathe and RestAC). Overall, the per-subject model performs better.



(a) Per-window (b) Per-subject

Figure 5.5.: Confusion matrices for the per-window and the per-subject approach. The true/actual labels are shown on the vertical axis and the predicted labels are on the horizontal axis. The biggest difference can be seen in the predictions of the Rest and Activity class.

5.2.3. DEEP LEARNING WITH HANDCRAFTED FEATURES

Current research mostly focuses on deep-learning networks for feature extraction and classification. Especially for heart rate variability analysis, there exist some standard features for measuring the variability. Although they are typically manually constructed, and therefore often interpretable, one may wonder if these measures capture all information needed for health diagnosis or activity recognition.

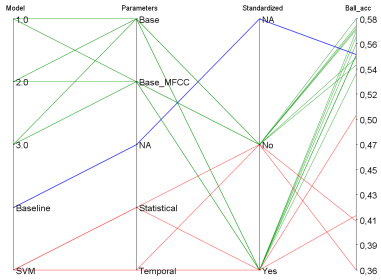
In this section, the statistical and temporal features used in earlier experiments are incorporated into a Deep Learning method to investigate whether incuding these handcrafted (HC) features can improve activity classification.
First of all, we compared the performance of the SVM models with the deep learning models (convolutional neural networks) with and without HC features. The results can be found in Figure 5.6 and illustrates that the addition of HC features results in an increase in balanced accuracies in comparison to the DL baseline model in certain instances. Additionally, the top four accuracies are achieved without standardizing the HC features. Furthermore, every DL model outperformed the SVM. Lastly, the second DL model which makes use of HC features and a window size of 80 and stride of 10, performs the best compared to all the models and configurations.

Next, we investigated the usage of temporal and statistical features as an alternative to the base set of HC features. Each DL model was trained with temporal or statistical features and with or without standardization. The outcomes can be found in Figure 5.6b. Among the top eight accuracies, five configurations employed standardized input. While in the previous experiment with the base features, the non-standardized HC features performed better. In addition to this, we combined both the statistical and temporal features into a single feature set, resulting in a slight improvement in performance to 58.84 % accuracy. Similarly, in this experiment, the standardized HC feature set worked better than the non-standardized one.

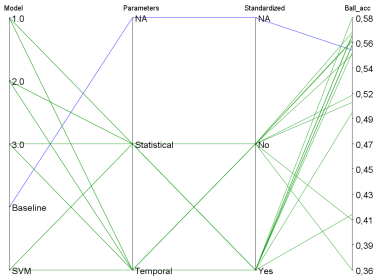
Besides solely examining the accuracies, it is relevant to investigate whether the HC features were indeed utilized by the DL model. To this extent, we used SHAP values to see how important the HC features are in addition to the raw input data [12]. Figure 5.7 depicts the top 20 SHAP values with the highest importance. As we can see, the highest SHAP values correspond to an HC feature: 0_Autocorrelation. Another observation is that primarily the heart rate values in the middle or the end of a window input are important (using a window size of 50).

5.2.4. MISCLASSIFICATION WITH DL MODELS

To interpret the predictions of the DL model with HC statistical and temporal features, a time series is plotted where the line colour indicates if the prediction is correct or incorrect, and to what class it is misclassified. An example is shown in Figure 5.8. Most of the misclassifications happen after a change of class (for instance, at t=270, where the class changes from Breathe to Activity). Intuitively this makes sense as the heart rate measurements do not immediately change during activity change and therefore it is difficult to predict the activity accurately.



(a) DL model base and MFCC features vs SVM statistical and temporal features.



(b) DL vs SVM, both using statistical and temporal features.

Figure 5.6.: Accuracies of different DL models with base HC features compared to DL baseline and SVM. The blue line is the baseline DL network that gets as input the raw standardized data. The red lines are the runs with an SVM and the green lines are the runs with the proposed DL networks. It can be seen that all the DL models outperform the SVM and most DL models with HC features outperform the baseline. The best performing configuration was achieved using model 2. Models 1, 2, and 3 represent the three different DL models, which are explained in the method section C.2. Models 1 and 3 make use of late integration and model 2 of early integration of the DL and HC features. The parameters column represents which feature set is used. The base feature set represents the basic HC features like max, min, mean, std and means of different (first and second-order) derivatives on the heart rate values in a time series window. Base and MFCC [10] represent the feature set where there are all the base parameters plus MFCC features. Statistical and temporal features are the features generated by TSFEL [11]. The column standardized indicates if the HC features are calculated on the standardized input or not. The raw heart rate time series data is always standardized. When we talk about standardized or non-standardized HC features in the next sections, we mean the features calculated on a standardized or non-standardized input.

5.3. DISCUSSION & CONCLUSION

We looked into the classification of activity using only heart rate time series. Results show that there seems to be a relation between the optimal window and stride size for classification: Higher window sizes and smaller strides correspond to higher accuracies. In this case, larger

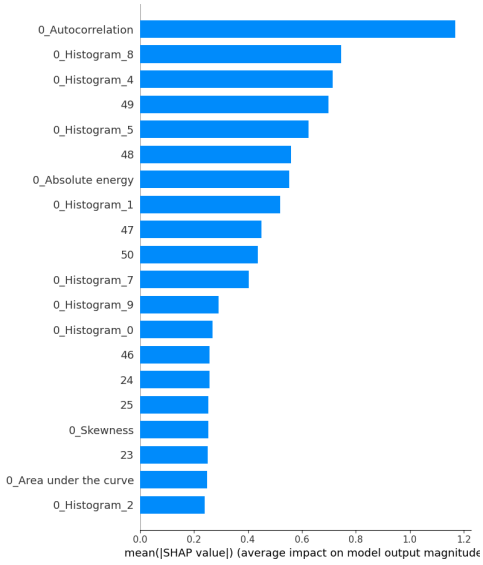


Figure 5.7.: Resulting top 20 SHAP values of 3rd DL model with the addition of temporal + statistical features. On the x-axis the SHAP value is shown and on the y-axis the feature. Features starting with '0_' indicate HC features, and numbers indicate the index of a heart rate value in the time series window input.

window sizes reveal distinct activity patterns that aren't discernible at smaller sizes. The smaller stride sizes ensure that the model is exposed to many time-shifted variations of an activity pattern, during training. This ensures that the model is better prepared to recognize any such variations that might be present during testing.
Because of the large variety in characteristics of wearables and persons (inter-device and inter-subject variability), it is challenging to build one single prediction model that works on everyone.
We have shown that in the context of activity classification in heart rate time series, it is helpful to group similar subjects together and thus creating semi-personalised models, for each separate group of alike persons. This improvement is particularly evident when using multiple windows from a subject to assign the person to a specific group. We believe that this is important to take into consideration when investigating more challenging tasks like heart disease classification.
Furthermore, we have shown that feeding a deep learning model with handcrafted features improves performance when classifying heart rate time series. By adding features manually, the network learns to find patterns in the time series itself but also uses some of the given HC

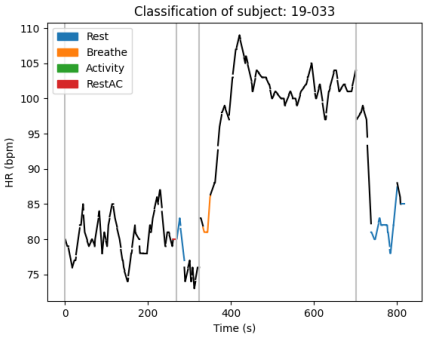


Figure 5.8.: Plot which visualizes the misclassification of the DL model. Black lines correspond to rightly predicted classes and a coloured section indicates a misprediction of a specific class. The colour corresponds to the class that is mispredicted. The grey vertical bars correspond to the change in activity according to the labeling of the data.

features to make a decision. This result supports the finding by Eltras et al.[13], where they use a DL network for feature extraction and concatenate the learned features with handcrafted features. Here, the best performance was achieved by concatenating the raw features with handcrafted features before inputting to the DL network.
Due to the limited number of subjects available, we did not have enough subjects in some situations to do a proper train/test split within a cluster, which was a common limitation for all the points mentioned above. This is particularly true for the the semi-personalised models. For those that were based on enough subjects, a positive increase in performance was shown. Another limitation of our work is that the experiments have only been conducted with one dataset. Finally, other normalization techniques can be explored to even further reduce the variability among the subjects, for example incorporating (meta)data like age or fitness.

REFERENCES

[1] J. Torres-Soto and E. A. Ashley. “Multi-task deep learning for cardiac rhythm detection in wearable devices”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–8. issn: 2398-6352.

[2] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G. H. Tison, G. M. Marcus, J. M. Sanchez, C. Maguire, and J. E. Olgin. “DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

[3] A. J. Dahalan, T. R. Razak, M. H. Ismail, S. S. M. Fauzi, and R. A. J. Gining. “Heart rate events classification via explainable fuzzy logic systems”. In: *IAES International Journal of Artificial Intelligence* 10.4 (2021), p. 1036. issn: 2089-4872.

[4] O. Walch, Y. Huang, D. Forger, and C. Goldstein. “Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device”. In: *Sleep* 42.12 (2019), zsz180. issn: 0161-8105.

[5] B. Bent and J. Dunn. *“BigIdeasLab_STEP”: Heart rate measurements captured by smartwatches for differing skin tones*" *(version 1.0)*. https://doi.org/10.13026/cqfy-d860.

[6] D. Maguire and R. Frisby. “Comparison of feature classification algorithm for activity recognition based on accelerometer and heart rate data”. In: *9th. IT & T Conference*, p. 11.

[7] J. Niu, Y. Tang, Z. Sun, and W. Zhang. “Inter-patient ECG classification with symbolic representations and multi-perspective convolutional neural networks”. In: *IEEE journal of biomedical and health informatics* 24.5 (2019), pp. 1321–1332. issn: 2168-2194.

[8] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. “Tslearn, A Machine Learning Toolkit for Time Series Data”. In: *Journal of Machine Learning Research* 21.118 (2020), pp. 1–6. url: http://jmlr.org/papers/v21/20-091.html.

[9] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa. “TSFEL: Time series feature extraction library”. In: *SoftwareX* 11 (2020), p. 100456.

[10] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian. “HMM-based audio keyword generation”. In: *Advances in Multimedia Information Processing-PCM 2004: 5th Pacific Rim Conference on Multimedia, Tokyo, Japan, November 30-December 3, 2004. Proceedings, Part III 5*. Springer, pp. 566–574. isbn: 3540239855.

[11] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa. “TSFEL: Time series feature extraction library”. In: *SoftwareX* 11 (2020), p. 100456. issn: 2352-7110.

[12] S. M. Lundberg and S.-I. Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[13] A. S. Eltrass, M. B. Tayel, and A. I. Ammar. “Automated ECG multi-class classification system based on combining deep learning features with HRV and ECG measures”. In: *Neural Computing and Applications* 34.11 (2022), pp. 8755–8775. issn: 1433-3058.

6

TACKLING INTER-SUBJECT VARIABILITY IN SMARTWATCH DATA USING FACTORIZATION MODELS

Smartwatches enable longitudinal and continuous data acquisition. This has the potential to remotely monitor (changes) of the health of users. However, differences among subjects (inter-subject variability) limit a model to generalize to unseen subjects. This study focused on binary classification tasks using heart rate and step counter from smartwatches, including night/day and inactive/active classification, as well as sleep and SpO2-related (oxygen saturation) tasks. To address inter-subject variability, we explored different transforming and normalization regimes for time series including per-subject and population-based strategies. We propose a modified Factorized Autoencoder (FAE), which separates the data into two latent spaces capturing class-specific and subject-specific information. Our proposed generalized factorized autoencoder (GFAE) and triplet factorized autoencoder (TFAE) improved classification accuracy over the baseline from 74.8 (±10.5) to 83.1 (±5.1) and 83.4 (±5.3), respectively, for night/day classification, gains for inactive/active classification were modest, improving from 84.3 (±9.4) to 86.9 (±4.4) and 86.6 (±4.3), respectively. Our study highlights challenges of handling inter-subject variability in smartwatch data and how factorization models. can be used to enable more robust and personalized health monitoring solutions for diverse populations.

This chapter has been published in Scientific Reports **15**, 26704 (2025).

6.1. INTRODUCTION

Consumer-grade smartwatches have transformed the landscape of personal health monitoring by enabling the continuous recording of vital signals, such as heart rate, physical activity, and sleep patterns. Originally developed for recreational use and fitness tracking, these devices have rapidly gained attention in the medical domain due to their widespread availability, non-invasiveness, and relatively low cost. When combined with machine learning (ML) techniques, smartwatch data has demonstrated significant potential for early detection and monitoring of various health conditions, including cardiovascular disease, diabetes, and respiratory disorders [1][2]. This shift marks an important step toward integrating wearable technologies into predictive and preventive healthcare.

In the context of cardiovascular disease, early detection and monitoring are particularly challenging due to the subtle, gradual changes in cardiovascular signals that precede acute events [3]. Cardiovascular disease affects millions worldwide, with significant mortality and morbidity rates, and is a leading cause of death [4][5]. Traditional diagnostic methods rely on echocardiography and clinical biomarkers[6][7], which are often inaccessible for continuous monitoring. Wearable devices provide an opportunity to bridge this gap by enabling remote, real-time monitoring of early warning signs such as changes in heart rate dynamics, physical activity, and sleep patterns. However, the successful implementation of such systems requires overcoming challenges related to variability and data quality.

Cardiovascular properties, including heart rate dynamics, are highly individualized and influenced by various factors. Studies have shown that body weight[8], sex[9], age, and height[10][11] play significant roles. Correspondingly, heart size and function scale with body mass index and exhibit sex-specific differences in cardiac output and structure. Furthermore, longitudinal studies using consumer-grade wearables like Fitbit have shown that resting heart rate and heart rate variability exhibit substantial inter-subject (between individuals) and intra-subject (within individuals over time) variability [12]. Such variability is further influenced by lifestyle factors, environmental conditions, and emotional states, complicating the interpretation of physiological signals for population-wide predictions.

Despite limitations, such as low sampling rates and reliance on derived metrics (e.g., estimated heart rate variability), smartwatches provide surprisingly detailed insights into physiological states. For example, studies have demonstrated their ability to capture heart rate patterns that correlate with respiratory infections [13] or cardiac conditions like atrial fibrillation and heart failure [14][15][16]. However, this level of detail also introduces challenges in ML-based analyses. Measurements of heart rate often encode individual-specific characteristics, which

may dominate the patterns learned by ML models, leading to poor generalization across unseen subjects. This issue arises when subject-specific variability (e.g., baseline heart rate differences) confounds the relationship between the physiological signal and the predictive task. Such confounding can degrade the performance of ML models, as they may focus on irrelevant patterns rather than generalizable features predictive of health outcomes.

Addressing inter- and intra-subject variability in smartwatch data has been a major focus in recent literature. Techniques such as domain adaptation[17], domain generalization[18], batch effect correction[19], and feature normalization[20] have been explored to improve generalization. Domain invariance techniques, including adversarial training and representation learning, aim to disentangle subject-specific factors from class-relevant signals[21][22][23]. These strategies are essential for creating robust predictive models that can generalize across diverse populations, a critical requirement for clinical applications of wearable data.

In this study, we propose a novel machine learning framework for predicting several tasks such as night/day and active/inactive binary classification, based on heart rate and step counter data obtained from smartwatches. Such tasks can for example help in assessing circadian rhythm disruptions or monitoring recovery post-surgery, which are important indicators of overall health. Our approach incorporates methods to address inter- and intra-subject variability, including normalization techniques, domain-invariant feature extraction and an extended loss function, to enhance the model's generalizability. By leveraging the continuous and rich data captured by consumer-grade wearables, we aim to provide a scalable solution for smartwatch monitoring, advancing the integration of wearable technology into cardiovascular healthcare.

6.2. METHODS

A general overview of our machine learning approach is given in Figure 1. Roughly, it consists of three steps. First, the smartwatch time series data is transformed to capture the dynamics and segmented into windows of several minutes. Then the data is normalized for which we propose and analyze several approaches (either population-based or subject-based). Lastly, the windows are fed into a factorized autoencoder that is specifically designed to disentangle variations between subjects and variations between classes by mapping the samples simultaneously in a so-called domain-space (representing differences between subjects) as well as a class-space (representing differences between the classes). For that, we do propose three different variants based on a contrastive technique, two of which are use our generalized factorized loss and one is based on our proposed triplet factorized loss. The following gives more details on each of the steps.

6.2.1. TRANSFORMATION

We apply two transformations to the heart rate time series to capture dynamic changes while minimizing the influence of individual baseline variability. Heart acceleration ($acc(t)$) is calculated as the first derivative of the heart rate hr, defined as the difference between the heart rate at time t and $t-1$ (Eq. 6.1). This metric isolates the temporal rate of change in heart rate, making it insensitive to baseline heart rate differences, which vary significantly across subjects and may confound comparisons. Additionally, we compute the normalized (relative) heart acceleration ($acc_{rel}(t)$) as the ratio of heart acceleration to the heart rate at time t (Eq. 6.2). This normalization accounts for variability in heart rate magnitude, enabling more consistent interpretation of changes across different heart rate levels:

$$acc(t) = hr(t) - hr(t-1) \tag{6.1}$$

$$acc_{rel}(t) = \frac{hr(t) - hr(t-1)}{hr(t)} \tag{6.2}$$

6.2.2. SEGMENTATION

The (transformed) heart rate timeseries is segmented using a sliding window of 240 samples (20 minutes) with a stride of 36 samples (3 minutes), at a sampling rate of once per 5 seconds. This window length was selected after hyperparameter tuning which included windows of length 120 (10 minutes), 360 (30 minutes) and 720 (60 minutes). To ensure data quality, windows containing timegaps of larger than 5 seconds, indicative of missing values are excluded. Each window is then assigned a label according to the criteria in Table 6.1. An SpO2 lower

Table 6.1.: Tasks (rows) and corresponding derived class labels (columns).

Task	Class 1	Class 2
Night/day	timestamp between 00:00 - 05:00	timestamp between 11:00 - 18:00
Inactive/active	Average steps <5 per minute	Average steps >50 per minute
SpO2 (ab)normal	Average SpO2 <90	Average SpO2 >95
Light/deep sleep	light sleep	deep sleep
Light/restless sleep	light sleep	restless sleep
Light/REM sleep	light sleep	REM sleep

than 95% is generally undesirable[24], but as a margin to distinguish abnormal events from normal ones, we label SpO2 levels lower than 90% as abnormal.

6.2.3. NORMALIZATION

We aim to further reduce inter-subject variability in heart rate data through the application of multiple techniques based on Z-normalization. Our approach leverages multiple variations of Z-score normalization tailored to different subsets of the data to enhance the consistency and comparability of heart rate measurements across individuals while accounting for subject-specific characteristics.

For population-based normalization, the heart rate time series of all training samples across all training subjects are aggregated to compute the population mean and standard deviation. The heart rate values are then standardized by subtracting the population mean and scaling them to unit variance. The calculated population mean and standard deviation are subsequently applied to normalize the heart rate time series of test subjects.

For per-subject normalization, Z-score normalization is applied individually to each subject. Two approaches are considered (and illustrated in Figure 6.1B): In the *per-test subject normalization*, after training with the population-based normalization, an initial portion of each test subject's data—up to the first 60% of their time series-is used to calculate a subject-specific mean and variance. The remaining test data is then normalized using these parameters. This approach ensures that each subject's data is standardized based on their unique heart rate characteristics. However, it requires a "burn-in" period to gather sufficient data for effective calibration. Note that this only affects the

model during the test phase. For the *per-train subject normalization* the Z-score normalization is performed separately for each training subject, using their individual mean and variance, computed on the entire training data of that subject. Note that this only affects the model during the training phase. The test subjects are processed similarly to the per-test subject scenario.

In both the population-based and per-subject normalization approaches, we explore two variations: normalizing based on all heart rate samples or exclusively on heart rate samples corresponding to zero steps (inactivity normalization). The motivation for the latter is to better estimate resting heart rate, as it focuses on periods of inactivity, which are less influenced by diverse and subject-specific physical activities. Normalization during inactive moments is hypothesized to provide a more consistent baseline across subjects, thereby enhancing the effectiveness of the normalization process.

6.2.4. FACTORIZED AUTOENCODERS

We investigated methods to mitigate inter-subject variability using contrastive[25] and triplet[26] similarity learning. Specifically, we adopted the Sequential Factorized Autoencoder (FAE) framework used by Gyawali et al.[21]. This approach addressed inter-subject variability by employing two key components. *Factorized latent representations*: The bottleneck layer of an autoencoder is partitioned into two distinct latent spaces: a class latent space (z^c) and a domain latent space (z^d). *Contrastive loss function*: A contrastive loss is applied to encourage the network to optimize class separability within the class latent space while making it invariant to subject-specific variability. Simultaneously, the domain latent space is trained to capture subject-specific characteristics while remaining independent of class labels. This separation allows the class latent space representation to be effectively used for downstream tasks.

Building on this framework, we generalized the original loss function and propose the Generalized Factorized Autoencoder (GFAE) defined as:

$$L = L^{ce} + \alpha\left(\beta L^c(z_i^c, z_j^c, y_{ij}^c) + (1-\beta)L^d(z_i^d, z_j^d, y_{ij}^d)\right) + \gamma L^{rec} \tag{6.3}$$

Where L^c is the class loss defined by the similarity between a pair of class latent samples (z_i^c and z_j^c) and the pair's corresponding class label y_{ij}^c. Similarly, L^d is the domain loss between a pair of domain latent samples (z_i^d and z_j^d) and the pair's corresponding domain label y_{ij}^d, where β trades off the class loss and domain loss. L^{ce} and L^{rec} represent

the cross-entropy loss and reconstruction loss and are defined as:

$$L^{ce} = -\frac{1}{2} \sum_{l \in \{i,j\}} \left(y_l^c log(\hat{y}_l^c) + (1 - y_l^c) log(1 - \hat{y}_l^c) \right) \tag{6.4}$$

$$L^{rec} = \frac{1}{2} \sum_{l \in \{i,j\}} ||x_l - \hat{x}_l||^2 \tag{6.5}$$

α and γ are the regularization coefficients for the contrastive loss and reconstruction loss, respectively.

The class loss is defined as:

$$L^c(z_i^c, z_j^c, y_{ij}^c) = y_{ij}^c ||z_i^c - z_j^c||_2^2 + (1 - y_{ij}^c) max(0, m^c - ||z_i^c - z_j^c||_2^2) \tag{6.6}$$

where y_{ij}^c equals one if the samples that form the pair have the same class labels, and zero otherwise. Regardless of which subjects the pair originates from, the first term aims to project samples of the same class close to each other in the class latent space. When the samples do not have the same class labels, they are kept distant of each other up to a threshold defined by the class margin, m^c.

The domain loss is similarly defined as:

$$L^d(z_i^d, z_j^d, y_{ij}^d) = y_{ij}^d ||z_i^d - z_j^d||_2^2 + (1 - y_{ij}^d) max(0, m^d - ||z_i^d - z_j^d||_2^2) \tag{6.7}$$

where y_{ij}^d equals one if the pair have the same domain labels and zero otherwise. Regardless of which classes the pair originates from, the first term aims to project samples of the same subject close to each other in the domain latent space. When the samples do not have the same domain labels, they are kept distant of each other up to a threshold defined by the domain margin, m^d.

In the original FAE, the parameter β was set to 0.5, equally balancing the class and domain losses. However, depending on the nature of the data, the model may benefit from adjusting this weighting to account for varying levels of inter-subject variability. In particular, the severity of inter-subject differences can influence how much emphasis should be placed on minimizing class vs. domain loss.

Additionally, in the original FAE framework, the domain loss did not account for situations where pairwise latent representations originated from different subjects. Specifically, in Equation 6.7, y_{ij}^d is always set to 1, effectively deactivating the second term. Explicitly modeling this scenario in the loss can enable the model to learn differences between subjects, improving subject separability. To address this limitation, we introduced the domain margin loss, adding it as the second term in equation 6.7. This modification is a critical consideration for

improving subject-specific invariance by explicitly modeling inter-subject differences.

Furthermore, the cross-entropy loss was not included in the original training process, as classification was performed during fine-tuning after the FAE model had been trained solely with the reconstruction loss. However, using reconstruction error as a surrogate loss function for model hyperparameter tuning did not yield models that performed well in classification tasks[27]. This suggests that optimizing solely for reconstruction error is insufficient for achieving high classification performance, emphasizing the need for a more targeted approach that integrates both class separability and domain invariance from the outset.

Next to the contrastive loss, we investigated using a triplet loss[26] instead, which we denote as the Triplet Factorized Autoencder (TFAE). It operates on triplets of samples: an anchor, a positive (same class as anchor) sample, and a negative (different class from anchor) sample. It learns to minimize the distance between the anchor and positive sample while maximizing the distance between the anchor and the negative sample.

The corresponding triplet class loss and domain loss are defined as:

$$L^c(z_i^d, z_j^d, z_k^d) = max\left(||z_i^c - z_j^c||_2^2 - ||z_i^c - z_k^c||_2^2 + m^c, 0\right) \tag{6.8}$$

$$L^d(z_i^d, z_j^d, z_k^d) = max\left(||z_i^d - z_j^d||_2^2 - ||z_i^d - z_k^d||_2^2 + m^d, 0\right) \tag{6.9}$$

It is important to note that sample i,j and k are the anchor, positive and negative sample, respectively. Therefore, the positive sample must have the same class and subject label as the anchor, while the negative sample must have a different class and subject label.

In the performed experiments, the factorized models are compared to a Multilayer Perceptron (MLP) baseline that is identical in number of neurons in each layer and all shared hyperparameters, to the factorized models except that its loss function includes only cross-entropy and reconstruction loss (Equation 6.4 and 6.5). Hyperparameter tuning using gridsearch resulted in the following optimal settings for the loss function: a value of 0.1 for α, 0.75 for β, 1.0 for γ and 0.01 for δ. For training, we used the Adam optimizer with a learning rate of 0.1, a batch size of 128 and a maximum of 100 epochs with early stopping. Finally, the optimal network architecture included a final hidden layer with 20 neurons. The MLP baseline and factorization models are based on a neural network consisting of in total 3 layers in the encoder where the number of neurons decay linearly. Thus, there are 240 neurons in the input layer, 130 neurons in the first hidden layer and 20 neurons in the second hidden layer. Furthermore, we inspect the separability of the domain and class latent space considering four scenarios: (1) a logistic regression trained on the domain latent space using the domain

(subject) label, which should be able to seperate subjects well in the domain space but not in the class space as that should not separate on subjects; (2) training on the domain latent space using the class label, which should not be able to separate classes both in the domain and class space as the domain space should not separate classes; (3), training on the class latent space using the label, which should be able to separate classes well in the class space but not in the domain space; (4) and training on the class latent space using the domain label, which should not be able to separate classes in both spaces. Only the latter two apply to the MLP, as it only learns one latent space.

6.2.5. EVALUATION

To evaluate the effectiveness of the proposed methods, we employ stratified leave-10-subjects-out cross-validation and use the ROCAUC score for the classification task as the performance metric. The data is split such that no subject appears simultaneously in both the training and validation sets, preventing information leakage. Furthermore, stratification is achieved through the pairwise and triplet sampling strategies (Appendix D.1), which ensure that each subject is sampled an equal number of times. Additionally, the pairwise sampling guarantee that the number of pairs sampled both within subjects and between subjects is balanced. Class labels are also sampled randomly to maintain an equal representation of both classes. Similarly, the triplet sampling approach ensures a balanced number of anchors from each class while also maintaining an equal distribution of samples from each subject.

6.3. RESULTS

6.3.1. DATA SET

Smartwatch data from the ME-TIME study (registered at clinicaltrials.gov with ID NCT05802563) was used and Table 6.2 show the characteristics of the development set, used to train and tune the models, as well as the test set. The diversity among subjects, is reflected in their corresponding smartwatch data. This is particularly evident in the mean and standard deviation of the heart rate, but also in the standard deviation of the (relative) heart acceleration, which exhibit substantial inter-subject variability, as shown in Figure 6.2. Such variability can significantly impact the generalization performance of machine learning models.

Furthermore, we have defined several binary classification tasks, as outlined in Table 6.1. The night/day classification is based on the timestamps of the heart rate sensor. Windows that fall entirely between midnight and 5 AM are labeled as ’night,’ while those that fall entirely between 11 AM and 6 PM are labeled as ’day.’ The inactive/active

classification is determined using the step counter, which records the number of steps taken per minute. Windows with an average of fewer than 5 steps per minute are classified as 'inactive,' allowing for minor step counts due to hand movement noise. Windows with more than 50 steps per minute are classified as 'active.' To maintain clear class separation, windows with step counts between 5 and 50 are excluded. The normal/abnormal SpO2 classification is based on blood oxygen saturation levels. Windows with SpO2 values below 90 are labeled 'abnormal,' while those above 95 are labeled 'normal.' Values between 90 and 95 are excluded to ensure a distinct separation between classes. Finally, the sleep stage classification is based on labels provided by the Fitbit. Only windows in which the sleep stage remains consistent throughout the entire window are considered.

Table 6.2.: Characteristics of the development and test sets.

Characteristic		**Train subjects**	**Test subjects**
Total		50	30
Age, years	18-39	17	9
	40-54	10	5
	55-64	9	4
	>65	14	12
Diagnosis	Reference	26	16
	Heart Failure	8	5
	Persistent atrial fibrillation	3	2
	Permanent atrial fibrillation	4	1
	Paroxysmal atrial fibrillation	9	6
Sex	Male	27	16
	Female	23	14
BMI	18.5–24.9	20	15
	25–29.9	16	9
	>30	14	6
Diabetes	Yes	19	5
	No	41	25
Smoking	Yes	13	4
	No	37	26
Hypertension	Yes	17	9
	No	33	21
Device	Charge 5	32	19
	Inspire 2	18	11

6.3.2. MULTI-SUBJECT FACTORIZATION

To evaluate the models' performance on multiple subjects, the models are trained using 50 train subjects and tested on 30 test subjects. The ROCAUC for the Night vs. Day classification (left) and Inactive vs. Active classification (right) is given in Table 6.3. The corresponding Receiver Operating Characteristic (ROC) curves for the unnormalized heart rate case are given in Figure D.6, Appendix D.5.

Table 6.3.: Mean and standard deviation of ROCAUC for night/day and inactive/active classification. For visual clarity, ROCAUC values are reported scaled from the standard range of 0–1 to 0–100. Hr=heart rate, hr,a=heart acceleration and hr,ar= relative heart acceleration. The GFAE is considered with a domain margin loss equal to zero ($m^d = 0$) and non-zero ($m^d \neq 0$), where after gridsearch, the optimal m^d was found to be 1.

Model/Task	Night & Day			Inactive & Active		
	hr	hr,a	hr,ar	hr	hr,a	hr,ar
MLP	74.81 (10.53)	73.24 (5.83)	58.63 (7.02)	84.12 (9.41)	82.11 (7.83)	75.63 (7.22)
FAE	78.23 (9.01)	71.12 (5.34)	70.63 (5.72)	85.91 (6.41)	81.72 (8.61)	75.13 (7.82)
GFAE ($m^d = 0$)	81.24 (5.31)	73.13 (6.31)	73.23 (6.01)	**87.83 (4.62)**	81.61 (9.13)	74.53 (8.84)
GFAE ($m^d \neq 0$)	83.14 (5.12)	74.31 (6.12)	74.12 (5.81)	86.93 (4.42)	81.41 (9.23)	74.71 (8.63)
TFAE	**83.42 (5.33)**	74.21 (6.23)	74.34 (5.61)	86.62 (4.31)	81.54 (9.11)	74.73 (8.41)

When using the original heart rate, the factorization models show improvement on the night/day classification task and to a lesser degree, improvement on the inactive/active task. The smaller improvement gain can be accounted to the fact that the latter is an easier task, with almost 10 higher points of ROCAUC (on a 0-100 scale) for the MLP baseline.

Using heart acceleration and relative heart acceleration normalizations does not improve the models. This may be due to the fact that heart rate values at adjacent time points are frequently identical, resulting in a one-point difference of zero. Consequently, the time series becomes sparse, as illustrated in Appendix D.4.

Furthermore, the TFAE performs best by a margin on the night/day classification and the GFAE without domain margin loss on the inactive/active classification. However, the differences are small

considering the fact that the GFAE and TFAE are within 0.5 points with a ten times higher standard deviation. Similarly, for the inactive/active classification task the GFAE and TFAE are within 1.5 points with a standard deviation of more than 4, indicating that the difference is relatively small. Overall, the standard deviation over subjects decreases when using factorization.

In Table 6.4, several additional tasks are considered. Both Light & REM sleep and Light & deep sleep tasks performs at near random for all models, while the normal & abnormal SPO2 performs slightly above random. The only task that performs better than random is classifying light sleep from restless sleep, where the GFAE with non-zero margin loss performs best. Furthermore, the GFAE in this configuration also has the best mean performance in three out of four tasks. Fitbit identifies restless sleep based on movement, such as tossing and turning, which notably impacts heart rate. In contrast, light and deep sleep have more subtle effects on heart rate that the Fitbit fails to capture [28].

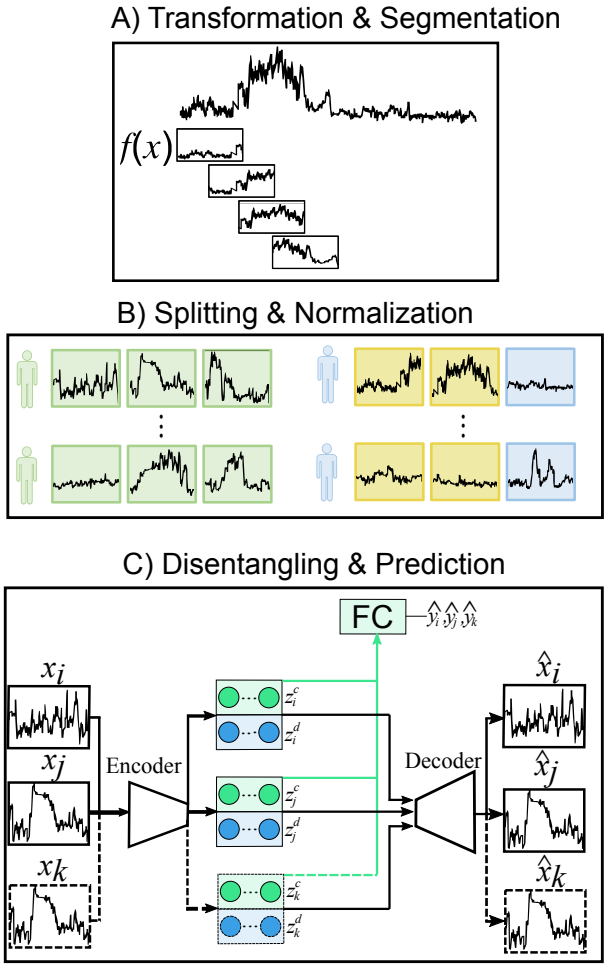


Figure 6.1.: Overview of the steps of the proposed machine learning approach. A) The timeseries signal is transformed and segmented using a sliding window. B) Subjects are split as train (green, left) or test (blue, right) subjects. Windows of test subjects are further split in calibration (yellow) and test (blue) windows. C) Windows are fed in factorized autoencoder models (three at a time for the triplet factorized autoencoder), where x_i, x_j refer to windows from the same subject with the same class label, and x_k refers to a window of a different subject and class label. The corresponding loss function consists of three main components (Equation 6.3). A fully-connected (FC) layer uses only the class latent space z^c to predict the class label to determine the cross-entropy loss. Both the domain latent space z^d and class latent space z^c are fed into the decoder to reconstruct the original input windows to determine the reconstruction loss. Finally, the class latent space and domain latent space are optimized using either our generalized factorized loss or triplet factorized loss, as described in Section 6.2.4.

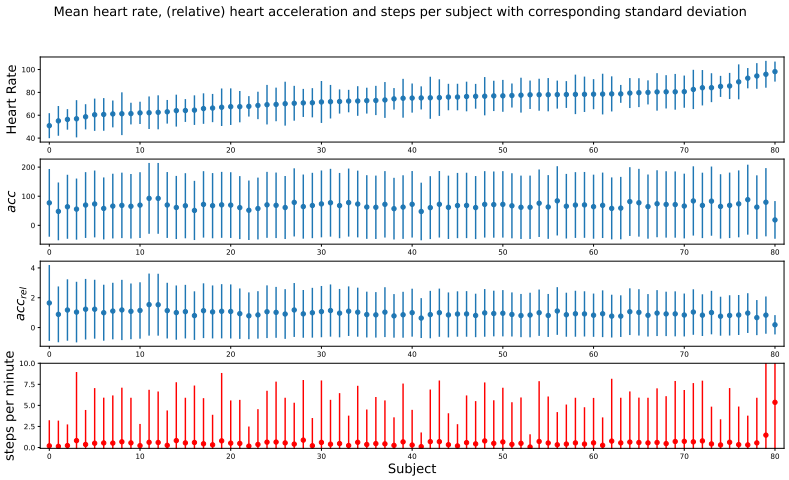


Figure 6.2.: Mean and standard deviation of data per subject for heart rate (first row), heart acceleration (second row), relative heart acceleration (third row) and steps per minute (last row).

Table 6.4.: Mean and standard deviation of ROCAUC over subjects for sleep related and SP02 tasks.For visual clarity, ROCAUC values are reported scaled from the standard range of 0–1 to 0–100. Hr=heart rate, hr,a=heart acceleration and hr,ar= relative heart acceleration. The GFAE is considered with a domain margin loss equal to zero and non-zero.

Model/Task	Light & Restless	Light & REM	SPO2 Normal & Abnormal			Light & Deep Sleep		
	hr	hr	hr	hr,a	hr,ar	hr	hr,a	hr,ar
MLP	61.41 (10.72)	49.63 (3.41)	51.24 (2.42)	50.31 (1.32)	49.83 (1.72)	**52.93** (5.12)	48.41 (3.74)	51.72 (2.03)
FAE	62.93 (9.71)	50.02 (0.10)	50.92 (2.31)	51.71 (2.42)	50.02 (1.10)	49.01 (3.24)	49.15 (0.62)	51.82 (3.93)
GFAE ($m^d = 0$)	63.54 (8.51)	49.82 (4.73)	52.51 (3.42)	50.43 (3.71)	50.14 (2.91)	52.42 (3.31)	50.10 (1.41)	50.11 (1.32)
GFAE ($m^d \neq 0$)	**63.72** (8.71)	**50.14** (4.42)	**52.91** (3.92)	50.03 (3.92)	50.41 (2.73)	52.23 (3.02)	50.31 (1.74)	49.93 (1.53)
TFAE	63.62 (8.42)	49.71 (4.53)	52.84 (3.72)	50.32 (3.74)	50.32 (2.51)	52.14 (3.11)	50.21 (1.64)	49.74 (1.42)

6.3.3. PER-SUBJECT CALIBRATION

We investigated the effects of the normalization methods described in Section 6.2.3. The results for Z-normalization applied per test subject, using varying percentages of each test subject's calibration set (yellow part of Figure 6.1b), are presented in Figure 6.3. Figure 6.4 illustrates the results of Z-normalization applied per train subject, where train subjects were normalized individually, in combination with per-test-subject normalization.

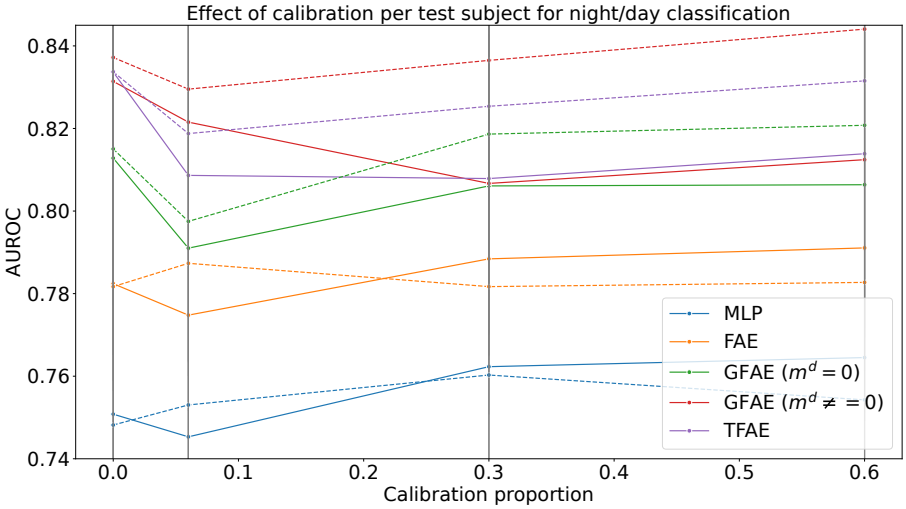


Figure 6.3.: Effect of calibration per-test-subject for night/day classification. Calibration proportion is the amount of data from a test subject used to compute a subject-specific mean and variance, to standardize the remaining data. Vertical grey lines denote the tested calibration proportions: 0%, 6%, 30%, and 60%. Dashed lines denote inactivity normalization.

Per-test subject normalization does not consistently improve performance. This is most pronounced in the GFAE with domain margin loss and the TFAE without inactivity normalization. Furthermore, normalization with too little calibration data can have adverse effects as can be seen by the dip in performance with a calibration proportion of 10%. Using all calibration data seems to benefit the mean performance for most models. In contrast, applying per-train subject normalization in addition to per-subject normalization provides a more consistent increase in performance with increasing calibration proportion.

Both configurations have been examined with and without the inclusion of inactivity normalization, which improves all models, albeit not consistently over all calibration proportions. The GFAE is the best performing model. When calibration is done per test subject, inactivity

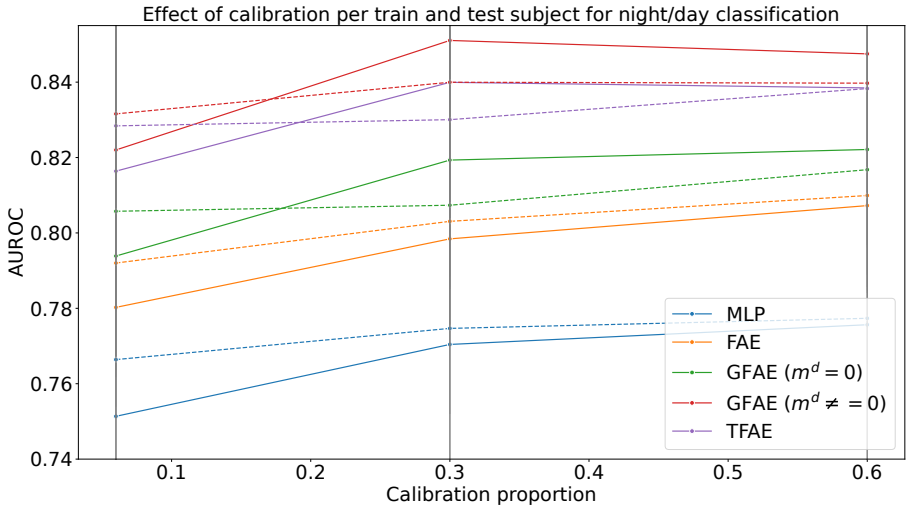


Figure 6.4.: Effect of calibration per-train and per-test-subject for night/day classification. Calibration proportion is the amount of data from a train or test subject used to compute a subject-specific mean and variance, to standardize the remaining data. Vertical grey lines denote the tested calibration proportions: 0%, 6%, 30%, and 60%. Dashed lines denote inactivity normalization. Since a population mean and standard deviation are not applicable in this configuration, it can only be performed with a non-zero calibration proportion.

normalization is required to reach a similar performance as calibration per train and test subject.

6.3.4. CLASS AND DOMAIN LATENT SPACE ANALYSIS

To quantify how well class and domain information is disentangled into their corresponding latent space, a logistic regression was trained on both latent spaces using either class labels, to inspect how well the classes are seperated (class accuracy) or domain labels, to inspect how well the subjects can be seperated (domain accuracy). Ideally, the class latent space should excel at class label prediction but perform poorly on subject labels, while the domain latent space should excel at subject label prediction but perform poorly on class labels. The train and test results for night day classification are presented in tables 6.5 and 6.6. Similarly, the results for inactive active classification can be found in Appendix D.3, tables D.1 and D.1.

Table 6.5.: Night/day class and domain train ROCAUC of logistic regression on z^c and z^d using 50 train subjects.

Model	Latent space	Class accuracy	Domain accuracy
MLP	z^c	89.71	8.52
FAE	z^c	82.42	10.21
	z^d	74.14	10.93
GFAE	z^c	90.52	14.11
($m^d = 0$)	z^d	87.70	16.24
GFAE	z^c	98.61	25.62
($m^d \neq 0$)	z^d	97.32	36.94
TFAE	z^c	98.43	10.23
	z^d	98.21	28.14

Table 6.6.: Night/day class and domain test ROCAUC of logistic regression on z^c and z^d using 30 test subjects.

Model	Latent space	Class accuracy	Domain accuracy
MLP	z^c	76.81	9.12
FAE	z^c	80.23	3.81
	z^d	76.12	4.53
GFAE	z^c	82.21	10.24
($m^d = 0$)	z^d	76.71	11.01
GFAE	z^c	83.02	12.22
($m^d \neq 0$)	z^d	73.10	14.70
TFAE	z^c	84.44	17.74
	z^d	74.31	20.11

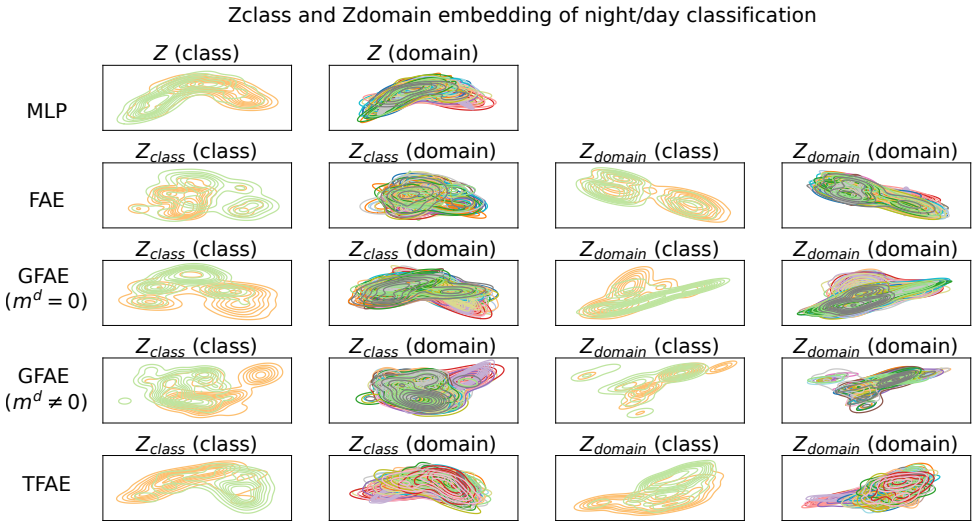


Figure 6.5.: UMAP visualization of the models using the test data for night/day classification. Parentheses indicate coloring (class or domain).

Interestingly, for the MLP model, the domain accuracy is slightly lower on the training set than on the test set, likely due to the increased complexity of the training set classification task. Nevertheless, all models, including the baseline MLP, surpass the 2% and 3.3% baseline accuracy for the train and test performance respectively (as a results of having 50 subjects in the train set and 30 subjects in the test set), demonstrating their ability to extract meaningful patterns despite the challenges posed by the subject distribution.

The results indicate further that both latent spaces contain substantial class information, as evidenced by their strong performance on the class labels. This can also be observed in the first and third column of the Uniform Manifold Approximation and Projection (UMAP) (Figure 6.5, D.3 and **??**), where the class and subject latent spaces are colored by class label. Still a significant class difference can be observed in z^d. Similarly, the domain latent space of factorization models often outperform the MLP in subject classification, suggesting a correlation between domain and class features, which could stem from the fact that the loss functions of the factorization models do not explicitly enforce feature decorrelation. although the factorization model tries to explicitely put the subject information on the domain latent space. Further analyses on pairwise subject classification (Appendix D.2), confirm this observation.

For both the train and test sets, the GFAE with positive domain margin loss as well as the TFAE improve domain accuracy using z^d compared

to the MLP. This demonstrates that the factorization models effectively capture and encode subject-specific information in the domain latent space. However, these improvements are less pronounced in the test set, particularly for the inactive/active task, suggesting that the factorization models could benefit from more subjects to generalize better.

6.4. DISCUSSION & CONCLUSION

We have investigated the challenge of inter-subject variability in heart rate time series, that limits the generalizability of machine learning models to unseen subjects. By evaluating normalization techniques and factorization models, we aimed to reduce this variability and improve classification performance for tasks such as night/day and inactive/active classification.

The heart rate time series that were transformed by the (relative) first-difference showed degraded performance, likely due to the sparsity it introduces in the signal, as heart rate values at adjacent time points are frequently identical, resulting in zeros or a one-point difference of zero. Factorized autoencoder models demonstrated consistent improvements, particularly for night/day classification. The GFAE with a positive domain margin loss as well as the TFAE achieved the highest accuracy, highlighting the benefit of explicitly modeling inter-subject variability in the loss function.

Using the factorization models as base, we calibrated them for each subject individually, using either Z-normalization on all the heart rate data or only during inactive moments. We investigated this calibration in two situations: calibrating only test subjects using a withheld calibration set or calibrating both train and test subjects. Combining normalization by the heart rate during inactive periods (referred to as inactivity normalization) with calibration across both train and test subjects yielded the most consistent improvements.

The factorization models were further investigated by using a logistic regression to investigate the separability of the classes in the class-related latent space and the subjects in the subject-related latent space.

The factorized models improved separability of subjects in the subject-related space. However, while the improvements over the baseline were significant, the separability still leaves substantial room for further refinement. Latent space analysis revealed that class-related information was encoded in both task and domain-latent spaces, potentially due to the loss functions' lack of explicit constraints to enforce feature decorrelation. Future work could explore techniques to enforce this by incorporating an adversarial component, as demonstrated in the sensor-based human activity recognition literature[29][30], could be

beneficial.

The addition of a domain margin loss allowed the models to learn differences between subjects, which is critical for addressing inter-subject variability. This loss explicitly models the distinctions between subjects in the subject-related latent space, enabling the model to better disentangle subject-specific characteristics. Future work could extend this by leveraging metadata on subjects, such as demographic and physiological information, directly in the domain margin loss.

Unfortunately, tasks such as sleep state and SpO2 classification, showed poor performance, where only the restless/light sleep task performed slightly better than random. The limited utility of these tasks is likely due to the lack of reliable labels from consumer-grade devices like Fitbit[31]. Furthermore, sleep stage is estimated using body movement data from the accelerometer and heart rate variability (HRV) derived from the PPG sensor—granular information that may not be fully preserved in the estimated heart rate in smartwatches.

By advancing methods to handle inter-subject variability, as demonstrated in this study, machine learning applications in smartwatch health monitoring will be better equipped to generalize across diverse populations. Our proposed generalized factorized autoencoder and triplet factorized autoencoder showed improvements using smartwatch data, highlighting their potential to address inter-subject variability. These advancements not only contribute to more accurate and reliable models but also pave the way for personalized and inclusive health monitoring solutions, ensuring greater applicability to real-world scenarios.

REFERENCES

[1] F. Sabry, T. Eltaras, W. Labda, K. Alzoubi, and Q. Malluhi. “Machine learning for healthcare wearable devices: the big picture”. In: *Journal of Healthcare Engineering* 2022.1 (2022), p. 4653923.

[2] H. S. Saad, J. F. Zaki, and M. M. Abdelsalam. “Employing of machine learning and wearable devices in healthcare system: tasks and challenges”. In: *Neural Computing and Applications* 36.29 (2024), pp. 17829–17849.

[3] M. Noitz, C. Mörtl, C. Böck, C. Mahringer, U. Bodenhofer, M. W. Dünser, and J. Meier. “Detection of Subtle ECG Changes Despite Superimposed Artifacts by Different Machine Learning Algorithms”. In: *Algorithms* 17.8 (2024), p. 360.

[4] World Health Organization. *Cardiovascular diseases (CVDs)*. Accessed: 2025-02-04. 2021. url: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[5] N. Townsend, D. Kazakiewicz, F. Lucy Wright, A. Timmis, R. Huculeci, A. Torbica, C. P. Gale, S. Achenbach, F. Weidinger, and P. Vardas. “Epidemiology of cardiovascular disease in Europe”. In: *Nature Reviews Cardiology* 19.2 (2022), pp. 133–143.

[6] T. A. McDonagh, L. Blue, A. L. Clark, U. Dahlström, I. Ekman, M. Lainscak, K. McDonald, M. Ryder, A. Strömberg, T. Jaarsma, *et al.* “European Society of Cardiology Heart Failure Association standards for delivering heart failure care”. In: *European journal of heart failure* 13.3 (2011), pp. 235–241.

[7] Mayo Clinic Staff. *Heart Disease: Diagnosis and Treatment*. Accessed: 2025-02-04. 2024. url: https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124.

[8] M. S. Lauer, K. M. Anderson, W. B. Kannel, and D. Levy. “The Impact of Obesity on Left Ventricular Mass and Geometry: The Framingham Heart Study”. In: *JAMA* 266.2 (July 1991), pp. 231–236. issn: 0098-7484. doi: 10.1001/jama.1991.03470020057032. eprint: https://jamanetwork.com/journals/jama/articlepdf/386462/jama_266_2_032.pdf. url: https://doi.org/10.1001/jama.1991.03470020057032.

[9] S. M. Ryan, A. L. Goldberger, S. M. Pincus, J. Mietus, and L. A. Lipsitz. “Gender- and age-related differences in heart rate dynamics: Are women more complex than men?” In: *Journal of the American College of Cardiology* 24.7 (1994), pp. 1700–1707. issn: 0735-1097. doi: https://doi.org/10.1016/0735-1097(94)90177-5. url: https://www.sciencedirect.com/science/article/pii/0735109794901775.

[10] A. Oberman, A. R. Myers, T. M. Karunas, and F. H. Epstein. “Heart size of adults in a natural population-Tecumseh, Michigan: Variation by sex, age, height, and weight”. In: *Circulation* 35.4 (1967), pp. 724–733.

[11] D. Liao, R. W. Barnes, L. E. Chambless, R. J. Simpson, P. Sorlie, G. Heiss, and The ARIC Investigators. “Age, race, and sex differences in autonomic cardiac function measured by spectral analysis of heart rate variability—The ARIC study”. In: *The American Journal of Cardiology* 76.12 (1995), pp. 906–912. issn: 0002-9149. doi: https://doi.org/10.1016/S0002-9149(99)80260-4. url: https://www.sciencedirect.com/science/article/pii/S0002914999802604.

[12] G. Quer, P. Gouda, M. Galarnyk, E. J. Topol, and S. R. Steinhubl. “Inter-and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults”. In: *Plos one* 15.2 (2020), e0227709.

[13] Y. Chen, D. She, Y. Guo, W. Chen, J. Li, D. Li, and L. Xie. “Smartwatch-based algorithm for early detection of pulmonary infection: Validation and performance evaluation”. In: *Digital Health* 10 (2024), p. 20552076241290684.

[14] G. H. Tison, J. M. Sanchez, B. Ballinger, A. Singh, J. E. Olgin, M. J. Pletcher, E. Vittinghoff, E. S. Lee, S. M. Fan, R. A. Gladstone, *et al.* “Passive detection of atrial fibrillation using a commercially available smartwatch”. In: *JAMA cardiology* 3.5 (2018), pp. 409–416.

[15] A. Naseri, D. M. Tax, M. Reinders, and I. van der Bilt. “Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data”. In: *Heliyon* 10.21 (2024).

[16] J. Wasserlauf, C. You, R. Patel, A. Valys, D. Albert, and R. Passman. “Smartwatch performance for the detection and quantification of atrial fibrillation”. In: *Circulation: Arrhythmia and Electrophysiology* 12.6 (2019), e006834.

[17] Z. Zhou, Y. Zhang, X. Yu, P. Yang, X.-Y. Li, J. Zhao, and H. Zhou. “Xhar: Deep domain adaptation for human activity recognition with smart devices”. In: *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE. 2020, pp. 1–9.

[18] T. Zhu, I. Afentakis, K. Li, R. Armiger, N. Hill, N. Oliver, and P. Georgiou. “Multi-Horizon Glucose Prediction Across Populations with Deep Domain Generalization”. In: *IEEE Journal of Biomedical and Health Informatics* (2024).

[19] X. Shen, R. Kellogg, D. J. Panyard, N. Bararpour, K. E. Castillo, B. Lee-McMullen, A. Delfarah, J. Ubellacker, S. Ahadi, Y. Rosenberg-Hasson, *et al.* “Multi-omics microsampling for the profiling of lifestyle-associated changes in health”. In: *Nature Biomedical Engineering* 8.1 (2024), pp. 11–29.

[20] C. S. Kumar, K. Ramachandran, and A. Kumar. “Vital sign normalisation for improving performance of multi-parameter patient monitors”. In: *Electronics Letters* 51.25 (2015), pp. 2089–2090.

[21] P. K. Gyawali, B. M. Horacek, J. L. Sapp, and L. Wang. “Sequential factorized autoencoder for localizing the origin of ventricular activation from 12-lead electrocardiograms”. In: *IEEE Transactions on Biomedical Engineering* 67.5 (2019), pp. 1505–1516.

[22] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao. “Learning disentangled semantic representation for domain adaptation”. In: *IJCAI: proceedings of the conference*. Vol. 2019. NIH Public Access. 2019, p. 2060.

[23] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling. “Diva: Domain invariant variational autoencoders”. In: *Medical Imaging with Deep Learning*. PMLR. 2020, pp. 322–348.

[24] J. G. Röttgering, A. M. de Man, T. C. Schuurs, E.-J. Wils, J. M. Daniels, J. G. van den Aardweg, A. R. Girbes, and Y. M. Smulders. “Determining a target SpO2 to maintain PaO2 within a physiological range”. In: *PLoS One* 16.5 (2021), e0250740.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[26] E. Hoffer and N. Ailon. “Deep metric learning using triplet network”. In: *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer. 2015, pp. 84–92.

[27] W. M. Kouw and M. Loog. “An introduction to domain adaptation and transfer learning”. In: *arXiv preprint arXiv:1812.11806* (2018).

[28] M. Tsunoda, T. Endo, S. Hashimoto, S. Honma, and K.-I. Honma. “Effects of light and sleep stages on heart rate variability in humans”. In: *Psychiatry and clinical neurosciences* 55.3 (2001), pp. 285–286.

[29] J. Su, Z. Wen, T. Lin, and Y. Guan. “Learning disentangled behaviour patterns for wearable-based human activity recognition”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.1 (2022), pp. 1–19.

[30] H. Qian, S. J. Pan, and C. Miao. “Latent independent excitation for generalizable sensor-based cross-person activity recognition”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 13. 2021, pp. 11921–11929.

[31] S. Haghayegh, S. Khoshnevis, M. H. Smolensky, K. R. Diller, and R. J. Castriotta. “Accuracy of wristband Fitbit models in assessing sleep: systematic review and meta-analysis”. In: *Journal of medical Internet research* 21.11 (2019), e16273.

7

DISCUSSION & CONCLUSION

Cardiovascular disease remains the leading cause of death and its prevalence is globally rising, pressuring hospitals, physicians, patients and therefore our society as a whole. This necessitates techniques and tools that enable timely detection and ideally prevention of cardiovascular disease. This dissertation set out to explore modern techniques to address this issue. Specifically, we set out to explore the feasibility, limitations, and methodological innovations necessary to enable machine learning–based cardiovascular monitoring using consumer-grade wearable devices.

7.1. FINDINGS

Using wearables for automatic detection of cardiovascular disease is fairly new. Therefore we performed a systematic review to assess the current state-of-the-art of using wearables (not limited by smartwatches) and machine learning for cardiovascular outcomes using the technology readiness level (TRL) characterisation in Chapter 2. Although most studies were performed in the last decade or so, most used the old MIT-BIH Arrhythmia database, which contains ECG recordings in the late 70's. Combined with the lack of cost-effective and non-burdensome devices offered by consumer-grade smartwatches and data acquisition in a strictly controlled environment, this jeopardizes the usability of the proposed models when deployed 'in-the-wild', which was also reflected through poor TRL scores of existing models (less than 6 out of a maximum score of 9). This underlines the need for using modern data sets from smartwatches from subjects in a free-living environment. To accommodate for this, we set up the ME-TIME study. An observational, longitudinal cohort study to acquire 'realistic' smartwatch data from subjects with atrial fibrillation (AF), heart failure (HF) along with healthy controls, acquiring data in a free-living, non-clinical setting (Chapter 2). Through initial experiments, we identified several important aspects and approaches for

the following methodological chapters.
Although there are a plethora of HRV measures that can be used (section 1.2), they are originally developed assuming the availability of instantaneous heart rate available at high sampling frequency. While some advanced smartwatches do provide access to high-frequency data, they cannot support continuous raw data transfer in a cost-effective way, due to battery life constraints. As a result, even with capable hardware, long-term high-frequency monitoring remains impractical, forcing a compromise in data acquisition. In practice, this means that consumer-grade smartwatches reduce sampling frequency by transmitting only estimated heart rate values — which further complicates the accurate estimation of traditional HRV measures. Therefore, there is a need for HRV measures that look for broader, zoomed-out patterns. To that end, we investigated the potential of new heart rate variability (HRV) measures to detect heart disease, namely heart rate - step counter cross-correlation maps (Chapter 2) and acceleration-deceleration curves (Chapter 3).
One of the recurring challenges throughout this dissertation is the scarcity of finely labeled data, especially when dealing with large-scale, real-world wearable datasets. In Chapter 3, we investigated the potential of solving this by applying weakly supervised learning techniques to the ME-TIME dataset, which includes timestamped heart rate and step count data but only a single diagnostic label per subject (i.e. the weak label). By leveraging multiple instance learning with embedded instance selection (MILES), we demonstrated that meaningful patterns associated with cardiovascular conditions can still be extracted under these weak-label conditions. Furthermore, in Chapter 3 we showed that an effective solution is to average the model outputs of acc-dec curves that fall in the same week and then use a final classification to classify the week. This is an important step toward making wearable-based diagnostics feasible at scale without excessive annotation demands.
Furthermore, we found that the heart rate and step counter show large variabilities between subjects (inter-subject variability), which can potentially harm the performance of machine learning models. Throughout the dissertation we showed different techniques to tackle this. We proposed normalizing the data of each subject using the step counter to reduce inter and intra subject variability. Furthermore in Chapter 4 we explored dividing subjects into groups and training a model on each group. During testing, we assessed to which group a test subject best fits, and applied the corresponding model.
Furthermore, we showed the effectiveness of using similarity learning through the triplet and contrastive loss to reduce inter-subject variability in several tasks in Chapter 5. The latter showed also to be effective in our heart disease.

7.2. IMPLEMENTATION CHALLENGES

A significant practical challenge in this study was the development and deployment of the infrastructure to collect, store, and manage wearable data via the Fitbit platform. Although the individual components required to build such a system are well-established, integrating them into a robust, scalable, and user-friendly research infrastructure proved unexpectedly difficult.

We specifically used Google Cloud Platform (GCP) as the hosting environment and even a minimal viable version of the system required familiarity and the orchestration of many services. BigQuery was used for structured data storage, Firestore to manage user authentication metadata, Cloud Run to execute serverless containers, which also run lacking services such as OAuth for authentication and API requests, and Cloud Scheduler to automate routine tasks. All components had to be wired together via custom Python scripts that handled the scheduling, API calls to Fitbit, token refresh logic, error handling, and data ingestion. This introduced considerable engineering overhead and debugging complexity, particularly given the absence of a unified orchestration layer.

Furthermore, there was no native support for handling multiple user accounts in a unified way. This was caused by the fact that all users required a Google account and the creation of these accounts were limited as a phone number must be used to verify the account and the same phone number could not be used many times for several accounts. As a workaround, all study participants had to be connected through a single, centralized Google account using Fitbit Device Connect, which significantly increased both development time and complexity. This workaround also introduced potential risks in terms of account throttling (there was a limit on number of API requests per hour per user) and token management (the risk of assigning the data of a patient to the wrong token and therefore the wrong patient).

Lastly, GCP's ecosystem lacked a suitable environment for rapid prototyping, debugging, and data inspection. As a result, it was necessary to export data from the cloud to a secure on-premise environment equipped with the required development tooling — such as an IDE, terminal access, and data visualization libraries. Moreover, the platform suffered from periods of instability, leading to corrupted data records or missing participant data, requiring manual scrutinizing and redoing the acquisition.

Beyond the technical issues, clinical integration was another bottleneck. Embedding this digital workflow into hospital procedures required protocol amendments to the medical ethical committee, staff training, and additional administrative coordination, which further delayed the adoption.

7.3. LIMITATIONS

Even though data were collected in free-living environments to enhance realism, the lack of additional contextual information about subjects limited the opportunity to get insights in free-living heart rhythm patterns. Such patterns arise due to the natural fluctuations in heart rate and rhythm that occur during everyday life including during physical activity, rest, consuming certain foods or beverages, emotional stress, or sleep. Identifying such variability could be used as inductive prior to the machine learning model. This contextual information could be captured, for example, by allowing participants to easily journal their daily routines and perceived well-being through a digital interface. In parallel, pathological smartwatch-derived HRV patterns within these contexts could be identified by a cardiologist by cross-referencing with simultaneously recorded ECGs. While the ultimate aim is to develop a fully autonomous cardiovascular monitoring system that does not rely on additional tools such as user input or ECG recordings, incorporating these modalities during the research phase can significantly enhance model development and interpretability.

7.4. FUTURE WORK

As technology continues to advance, the types of physiological sensors available for continuous and non-invasive cardiovascular monitoring is rapidly expanding. Current tools span a broad spectrum, from invasive devices such as implantable cardioverter-defibrillators (ICDs) to semi-invasive or clinical-grade monitors like chest straps, and increasingly, to consumer-grade wearables such as smartwatches. Looking ahead, emerging sensors such as flexible, wrist-based 'patchables' [1] hold promise for supplementing PPG sensors. On another branch of trends, different modalities are already being integrated into the smartwatch, such as electrodermic activity, temperature, SpO2 etc. Perhaps these two branches of trends will syncronize, where new continous, non-invasive wearables mature, until they can be integrated in a smartwatch.

However, for such systems to have real clinical impact a critical challenge is the scale and quality of available data. While this dissertation worked with real-world datasets, much larger and more representative data collections are required—both in terms of sample size and demographic diversity of patients (age, sex, physical activity levels, and socio-demographic context). Furthermore, substantial diversity in comorbidities and disease severity is essential to rigorously evaluate the robustness of machine learning models, particularly when the HRV signal indicative of heart disease is subtle during early disease stages, potentially enabling prognostic assessment. Future work should thus prioritize large-scale, multi-center, longitudinal studies. Such efforts will require coordinated infrastructure, possibly on a national or international level,

similar to biobank-scale initiatives but with a focus on wearable data and associated outcomes.

This scale of data collection brings both opportunities and challenges for ML development. Big data facilitates foundation models devised through self-supervised techniques, which subsequently enables the creation of smaller, less label-intensive downstream models. Nevertheless, population diversity complicates the generalization of ML models, making inter-subject variability correction critical. While richer annotated data through subject journaling can enhance models by providing essential context and enabling active learning methods, human-in-the-loop approaches introduce potential biases and burdens which must be carefully investigated in future work. For instance, subjects might (unintentionally) bias the type and moment of responses based on personal expectations, social desirability or psychological stress. Additionally, targeted journaling requests might inadvertently lead subjects to become overly concerned or anxious, misinterpreting these prompts as indicative of potential underlying health issues, thus potentially affecting both data accuracy and the individual's psychological well-being.

7.5. CLINICAL TRANSLATION

Implementing these advanced cardiovascular monitoring systems in clinical practice will substantially alter current operational paradigms. Healthcare providers must establish clear clinical guidelines and standards to ensure safe and effective usage of these technologies. Patients require education and clear communication to confidently interact with these monitoring systems without undue stress or misinterpretation. This involves designing intuitive interfaces and actionable alerts, coupled with robust clinical protocols to respond appropriately to early warnings or subtle physiological changes identified by the system. Regulatory bodies must address data privacy concerns, ethical implications, and validate the safety and efficacy of ML-driven insights before widespread adoption. Additionally, insurance companies should be regulated to ensure their policies only evolve in ways that genuinely benefit users, such as encouraging early preventive actions to reduce long-term medical costs rather than penalizing users based on cardiovascular assessments of the smartwatch. Future work should thus focus on multi-stakeholder collaborations to develop robust frameworks and regulations that support ethical deployment, ensure patient safety, and maintain data integrity and privacy.

7.6. CONCLUSION

This dissertation demonstrated the feasibility of using consumer-grade wearables for cardiovascular diagnosis through novel methodological approaches and real-world data collection. While challenges remain in data quality, clinical integration, and scalability, the findings provide a strong foundation for future research and system development. With continued innovation, wearable-based cardiovascular monitoring has the potential to meaningfully shift preventive care toward earlier, more personalized intervention.

REFERENCES

[1] S. Zhou, G. Park, K. Longardner, M. Lin, B. Qi, X. Yang, X. Gao, H. Huang, X. Chen, Y. Bian, *et al.* “Clinical validation of a wearable ultrasound sensor of blood pressure”. In: *Nature Biomedical Engineering* (2024), pp. 1–17.

[2] V. W. Anelli, T. Di Noia, E. Di Sciascio, C. Pomo, and A. Ragone. “On the discriminative power of hyper-parameters in cross-validation and how to choose them”. In: *Proceedings of the 13th ACM conference on recommender systems*. 2019, pp. 447–451.

[3] B. Bent and J. Dunn. *“BigIdeasLab_STEP”: Heart rate measurements captured by smartwatches for differing skin tones*" *(version 1.0)*. https://doi.org/10.13026/cqfy-d860.

[4] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. “A comprehensive survey on support vector machine classification: Applications, challenges and trends”. In: *Neurocomputing* 408 (2020), pp. 189–215.

[5] E. Mayoraz and E. Alpaydin. “Support vector machines for multiclass classification”. In: *Engineering Applications of Bio-Inspired Artificial Neural Networks: International Work-Conference on Artificial and Natural Neural Networks, IWANN’99 Alicante, Spain, June 2–4, 1999 Proceedings, Volume II*. Springer. 2006, pp. 833–842.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[7] K. G. Schmahl, T. J. Viering, S. Makrodimitris, A. Naseri, D. Tax, and M. Loog. “Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings”. In: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 2020, pp. 94–103.

[8] A. Naseri, D. Tax, M. Reinders, and I. van der Bilt. “Machine learning for cardiovascular outcomes from wearable data: systematic review from a technology readiness level point of view”. In: *JMIR medical informatics* 10.1 (2022), e29434.

[9] A. Naseri, D. Tax, P. van der Harst, M. Reinders, and I. van der Bilt. “Data-efficient machine learning methods in the ME-TIME study: Rationale and design of a longitudinal study to detect atrial fibrillation and heart failure from wearables”. In: *Cardiovascular Digital Health Journal* 4.6 (2023), pp. 165–172.

[10] M. Beekhuizen, A. Naseri, D. Tax, I. van der Bilt, and M. Reinders. “Improving performance of heart rate time series classification by grouping subjects”. In: *arXiv preprint arXiv:2311.13285* (2023).

[11] A. Naseri, D. M. Tax, M. Reinders, and I. van der Bilt. “Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data”. In: *Heliyon* 10.21 (2024).

[12] A. Naseri, D. Tax, I. van der Bilt, and M. Reinders. “Tackling inter-subject variability in smartwatch data using factorization models”. In: *Submitted to Scientific reports* (in review).

BIBLIOGRAPHY

[1] S. Zhou, G. Park, K. Longardner, M. Lin, B. Qi, X. Yang, X. Gao, H. Huang, X. Chen, Y. Bian, *et al.* “Clinical validation of a wearable ultrasound sensor of blood pressure”. In: *Nature Biomedical Engineering* (2024), pp. 1–17.

[2] V. W. Anelli, T. Di Noia, E. Di Sciascio, C. Pomo, and A. Ragone. “On the discriminative power of hyper-parameters in cross-validation and how to choose them”. In: *Proceedings of the 13th ACM conference on recommender systems*. 2019, pp. 447–451.

[3] B. Bent and J. Dunn. *“BigIdeasLab_STEP”: Heart rate measurements captured by smartwatches for differing skin tones*" *(version 1.0)*. https://doi.org/10.13026/cqfy-d860.

[4] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. “A comprehensive survey on support vector machine classification: Applications, challenges and trends”. In: *Neurocomputing* 408 (2020), pp. 189–215.

[5] E. Mayoraz and E. Alpaydin. “Support vector machines for multiclass classification”. In: *Engineering Applications of Bio-Inspired Artificial Neural Networks: International Work-Conference on Artificial and Natural Neural Networks, IWANN’99 Alicante, Spain, June 2–4, 1999 Proceedings, Volume II*. Springer. 2006, pp. 833–842.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[7] K. G. Schmahl, T. J. Viering, S. Makrodimitris, A. Naseri, D. Tax, and M. Loog. “Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings”. In: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 2020, pp. 94–103.

[8] A. Naseri, D. Tax, M. Reinders, and I. van der Bilt. “Machine learning for cardiovascular outcomes from wearable data: systematic review from a technology readiness level point of view”. In: *JMIR medical informatics* 10.1 (2022), e29434.

[9] A. Naseri, D. Tax, P. van der Harst, M. Reinders, and I. van der Bilt. “Data-efficient machine learning methods in the ME-TIME study: Rationale and design of a longitudinal study to detect atrial fibrillation and heart failure from wearables”. In: *Cardiovascular Digital Health Journal* 4.6 (2023), pp. 165–172.

[10] M. Beekhuizen, A. Naseri, D. Tax, I. van der Bilt, and M. Reinders. “Improving performance of heart rate time series classification by grouping subjects”. In: *arXiv preprint arXiv:2311.13285* (2023).

[11] A. Naseri, D. M. Tax, M. Reinders, and I. van der Bilt. “Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data”. In: *Heliyon* 10.21 (2024).

[12] A. Naseri, D. Tax, I. van der Bilt, and M. Reinders. “Tackling inter-subject variability in smartwatch data using factorization models”. In: *Submitted to Scientific reports* (in review).

A

APPENDIX OF CHAPTER 2

A.1. SEARCH QUERIES

Table A.1.: Search queries performed in the three electronic databases.

Database	Search query
Scopus	TITLE-ABS-KEY(("machine learning" OR "deep learning" OR "neural networks" OR "artificial intelligence") AND (wearable OR smartwatch OR "fitness tracker") AND (cardiovascular OR cardiology OR cardiac OR heart OR "atrial fibrillation" OR "heart failure" OR arrhythmia)) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO(PUBSTAGE,"final")) AND (LIMIT-TO(DOCTYPE,"ar")) AND (LIMIT-TO(LANGUAGE,"English"))
PubMed & IEEE Xplore	(("machine learning" OR "deep learning" OR "neural networks" OR "artificial intelligence") AND (wearable OR smartwatch OR "fitness tracker") AND (cardiovascular OR cardiology OR cardiac OR heart OR "atrial fibrillation" OR "heart failure" OR arrhythmia))

A.2. STUDY CHARACTERISTICS

Table A.2.: Benchmark, sample size, processing and TRL.

Study	Benchmark	Sample size	Processing	TRL
[1]		50	Computer	4
[2]	yes	124	Computer	5
[3]		401	Computer	4
[4]	Yes	104	Computer	5

A

Study	Benchmark	Sample size	Processing	TRL
[5]		30	Computer	4
[6]		501	Computer	4
[7]	Yes	330	Smartphone	5
[8]	Yes	104	Computer	5
[9]		30	Smartphone	4
[10]		85	Computer	4
[11]	Yes	NR	Computer	5
[12]		46	Computer	4
[13]		40	Computer	4
[14]		100	Computer	4
[15]		3777	Computer	5
[16]		75	Computer	4
[17]		49	Computer	4
[18]		97	Computer	4
[19]		70	Computer	4
[20]		100	Smartphone	4
[21]	Yes	487	Computer	5
[22]		83	Computer	5
[23]		368	Smartphone	5
[24]		220	Computer	5
[25]	Yes	52	Computer	5
[26]		45	Computer	4
[27]		8299	Computer	5
[28]		7524	Computer	5
[29]	Yes	43	Embedded	3
[30]	Yes	NR	Computer	3
[31]	Yes	44	Embedded	3
[32]	Yes	47	Computer	3
[33]	Yes	NR	Computer	3
[34]	Yes	47	Smartphone	3
[35]	Yes	NR	Smartphone	3
[36]	Yes	NR	Computer	3
[37]	Yes	181	Computer	3
[38]	Yes	NR	Computer	3
[39]	Yes	23	Computer	3
[40]	Yes	47	Computer	3
[41]	Yes	NR	Computer	3
[42]	Yes	47	Computer	3
[43]	Yes	47	Computer	3
[44]	Yes	43	Computer	3
[45]	Yes	47	Computer	3
[46]	Yes	43	Computer	3
[47]	Yes	43	Embedded	3

Study	Benchmark	Sample size	Processing	TRL
[48]	Yes	NR	Computer	3
[49]	Yes	NR	Computer	3
[50]	Yes	290	Computer	3
[51]	Yes	44	Computer	3
[52]	Yes	NR	Smartphone	3
[53]	Yes	104	Embedded	3
[54]	Yes	44	Computer	3
[55]	Yes	NR	Computer	3

Table A.3.: Study duration, observation period and recording duration. NR=Not Reported.

Study	Study Duration	Observation Period	Recording Duration
[1]	Mar 19 - Oct 2019	24H	24H
[2]	Feb 17 - Apr 18	24H	17H
[3]	May 19 - Jun 19	NR	3M
[4]	NR	NR	30M
[5]	Jun 7, 17	22M	22M
[6]	NR	NR	1M
[7]	NR	NR	NR
[8]	NR	NR	30M
[9]	NR	NR	NR
[10]	NR	NR	30S
[11]	NR	NR	5M
[12]	NR	14D	near continuous
[13]	NR	5.7M	5.7M
[14]	Aug 15 - Dec 16	90D	NR
[15]	train: Mar 13 - Jan 18	NR	median 10.5H
[16]	Sep 17 - Apr 18	30m	15M
[17]	NR	NR	NR
[18]	Oct 15 - Mar 16	5m	5M
[19]	Mar 15 - Nov 15	10m	10M
[20]	18 - 19	15m	15M
[21]	NR	NR	train: (1,2), test:60 R-R intervals
[22]	PIONEER-HCM:, Aug 16 - Nov 17 MYK-491-001: Jan016 - Nov 18	1D	5M
[23]	NR	NR	16S
[24]	NR	train: 45min,	NR

A

Study	Study Duration	Observation Period	Recording Duration
		test:1 week	
[25]	NR	NR	NR
[26]	NR	NR	6M
[27]	Feb 16 - Mar 17	NR	>8H per day
[28]	train: NR, test: May 17 - Sep 17	NR	11.3H
[29]	1975 - 1979	NR	30M
[30]	NR	NR	30M
[31]	1975 - 1979	NR	30M
[32]	1975 - 1979	NR	30M
[33]	NR	NR	10H
[34]	1975 - 1979	NR	30M
[35]	NR	NR	3H
[36]	NR	NR	30M
[37]	NR	NR	1M
[38]	NR	NR	1M
[39]	1975 - 1979	NR	10H
[40]	1975 - 1979	NR	30M
[41]	NR	NR	1M
[42]	1975 - 1979	NR	30M
[43]	1975 - 1979	NR	30M
[44]	1975 - 1979	NR	30M
[45]	1975 - 1979	NR	30M
[46]	1975 - 1979	NR	30M
[47]	1975 - 1979	NR	30M
[48]	NR	NR	2M
[49]	NR	NR	1M
[50]	NR	NR	2M
[51]	1975 - 1979	NR	30M
[52]	NR	NR	NR
[53]	NR	NR	NR
[54]	1975 - 1979	NR	30M
[55]	NR	NR	3H

Table A.4.: Activity, setting, placement and modality. C = Controlled, F = Free-living, A = Active, S = Sedentary.

Study	Environment	Setting	Placement	Modality
[1]	C	A	Chest	ECG
[2]	C	NR	Chest	ECG
[3]	C	S	wrist	PPG, ECG

Study	Environment	Setting	Placement	Modality
[4]	NR	NR	Textile	ECG
[5]	C	A	Chest, finger	ECG, Galvanic Skin Response
[6]	C	S	Handheld	PPG, ECG
[7]	C	A	Chest	ECG
[8]	NR	NR	Chest	ECG
[9]	NR	NR	Wrist	ECG
[10]	C	S	Wrist	ECG
[11]	NR	NR	NR	ECG
[12]	CF	A	Chest	ECG, accelerometer, location
[13]	C	S	Chest	seismocardiogram, gyro-cardiogram
[14]	F	A	Chest	ECG, accelerometer, skin impedance, temperature.
[15]	C	S	Finger	PPG
[16]	C	S	Finger	PPG
[17]	C	A	Head	near infrared spectroscopy
[18]	C	S	wrist	PPG, accelerometer
[19]	C	S	Wrist	PPG, accelerometer
[20]	C	S	Finger	PPG
[21]	C	S	Chest	ECG
[22]	C	S	Wrist	PPG
[23]	NR	NR	Chest	ECG
[24]	F	A	Wrist	PPG
[25]	F	A	Wrist	ECG
[26]	CF	A	chest	seismocardiogram
[27]	F	A	Wrist	PPG, accelerometer
[28]	F	A	Wrist	PPG, accelerometer
[29]	C	NR	Chest	ECG
[30]	C	NR	Chest	ECG
[31]	C	NR	Chest	ECG
[32]	C	NR	Chest	ECG
[33]	NR	NR	Chest, handheld	ECG
[34]	C	NR	Chest	ECG
[35]	NR	NR	Chest	ECG
[36]	NR	NR	Chest	ECG
[37]	NR	NR	Chest	ECG
[38]	NR	S	Handheld	ECG
[39]	C	NR	Chest	ECG

A

Study	Environment	Setting	Placement	Modality
[40]	C	NR	Chest	ECG
[41]	NR	S	Handheld	ECG
[42]	C	NR	Chest	ECG
[43]	C	NR	Chest	ECG
[44]	C	NR	Chest	ECG
[45]	C	NR	Chest	ECG
[46]	C	NR	Chest	ECG
[47]	C	NR	Chest	ECG
[48]	NR	NR	Chest	ECG
[49]	NR	S	Handheld	ECG
[50]	C	NR	Chest	ECG
[51]	C	NR	Chest	ECG
[52]	NR	NR	Chest	ECG
[53]	NR	NR	NR	ECG
[54]	C	NR	Chest	ECG
[55]	NR	NR	Chest	ECG

Table A.5.: Input window size, model, features and validation.

Study	Input window size	Model	Features	Validation
[1]	10S	CNN	WI	NR
[2]	variable, 3-beat units	CNN-RNN	R	GH
[3]	NR	Residual NN	R	GH
[4]	0.55S	CNN	R	GH
[5]	10S	MLP, logistic regression random forest	St, D	NR
[6]	variable, one beat	linear regression, decision tree, MLP	WI	GH
[7]	30S	CNN	R	CV
[8]	0.55S	Autoencoder	WI	CV
[9]	5S	Decision tree	WI	GH
[10]	30S	ensemble	St	CV
[11]	30S	LSTM	WI	GH
[12]	15S	Hierarchical Bayesian model	St, Sp	GCV
[13]	10S	Decision tree, random forest, MLP	St, Sp	CV
[14]	1M	Graph similarity	St, O	NR
[15]	30S	SVM, CNN	R	CV

Study	Input window size	Model	Features	Validation
[16]	30S	CNN, RNN, SVM	R	GCV
[17]	NR	MLP	O	GH
[18]	5m	SVM	St, D	CV
[19]	2m	kNN	St	GCV
[20]	30S	CNN	R	GCV
[21]	train: (1,2) test:60 RR intervals	SVM	St	CV
[22]	1 beat	MILES	WI	GCV
[23]	NR	MLP	R	GH
[24]	25S	CNN-autoencoder	R	H
[25]	2M	SVM, random forest	WI, St, Sp	GCV
[26]	1s	Graph similarity	Sp	NR
[27]	5s, 30s, 5m, 30m	CNN-LSTM	St, O	GH
[28]	1H	CNN	St	NR
[29]	2 RR intervals	MLP	WI	CV
[30]	0.24S	Echo state NN	WI	CV
[31]	1 RR interval	CNN-LSTM	R	H
[32]	0.22S	SVM	WI	GCV
[33]	30S	Gradient boosted tree	WI, St	CV
[34]	1 beat	MLP	WI	CV
[35]	2S	MLP	WI, St	H
[36]	25 RR intervals	MLP	WI	H
[37]	1M	Decision Table, MLP, random forest,kNN logistic regression, SVM, ensemble	WI	CV
[38]	1M	SVM, bagging trees	WI, Sp	CV
[39]	1S	CNN	R	H
[40]	1 P-QRS-T complex	SVM	R	CV
[41]	9S, 15S	CNN	Sp	SCV
[42]	0.7S around R-peak (0.25 left, 0.45 right)	spiking NN	R	H
[43]	0.8S around R-peak (0.4 left, 0.4 right)	CNN	R	SCV
[44]	0.5S around R-peak (0.25 left, 0.25 right)	spiking NN	R	GH
[45]	1.1S around R-peak (0.37 left, 0.74 right)	MLP	R	H
[46]	0.56S	LSTM	R	H
[47]	0.7S around R-peak (0.25 left, 0.45 right)	LSTM	R	H
[48]	1707ms	CNN-LSTM	R	CV
[49]	5S	CNN	R	CV

A

Study	Input window size	Model	Features	Validation
[50]	0.6S around R-peak (0.2 left, 0.4 right)	CNN-LSTM	R	CV
[51]	0.5S around R-peak	CNN	R	GH
[52]	20S	MLP	WI, St	CV
[53]	0.65S around R-peak (0.25 left, 0.4 right)	Random forest	WI, Sp	GCV
[54]	0.36S around R-peak	CNN	R	H
[55]	0.2S around R-peak (0.75 left, 0.125 right)	Echo state NN	WI	CV

Table A.6.: Type of cardiovascular disease used as target condition. AA=Atrial Arrhythmia, VA=Ventricular Arrhythmia, C=Control, SR=Sinus Rhythm, HF=Heart Failure, CP=Cardiovascular Prevention, CAD=Coronary Artery Disease, VHD=Valvular Heart Disease.

Study	Clinical outcome
[1]	AA, C
[2]	SR, AA, VA, C
[3]	AA, C
[4]	SR, AA, VA, C
[5]	SR, CP
[6]	SR, CP
[7]	SR, AA, VA, C
[8]	AA, VA
[9]	SR, CAD
[10]	SR, AA
[11]	SR, C
[12]	CP
[13]	VHD
[14]	HF
[15]	AA, C
[16]	SR, AA
[17]	CAD
[18]	HF, C
[19]	SR, AA, C
[20]	SR, AA
[21]	AA, C
[22]	SR, HF
[23]	SR, C
[24]	SR, AA

Study	Clinical outcome
[25]	AA, VA
[26]	HF
[27]	AA, C
[28]	AA, C
[29]	SR, AA, VA, C
[30]	SR, AA, VA, C
[31]	SR, AA, VA, C
[32]	AA, VA
[33]	SR, AA, C
[34]	SR, VA, C
[35]	SR, AA, VA
[36]	SR, AF
[37]	SR, AA
[38]	SR, AA, C
[39]	SR, AA, C
[40]	SR, AA, VA, C
[41]	SR, AA, C
[42]	SR, VA, C
[43]	SR, VA, C
[44]	SR, AA, VA, C
[45]	SR, AA, VA, C
[46]	SR, AA, VA
[47]	SR, AA, VA, C
[48]	SR, CAD, C
[49]	SR, AA, C
[50]	SR, CAD
[51]	SR, AA, VA, C
[52]	VA, C
[53]	SR, CAD
[54]	AA, VA
[55]	VA, C

B

APPENDIX OF CHAPTER 4

B.1. VALIDATION AND HYPERPARAMETER TUNING

Utilizing gridsearch and Stratified Leave-6-Subjects-Out Cross-Validation (SL6SOCV) on the development set, we fine-tuned the model's hyperparameters[2]. Stratification maintains an equal number of Ref and CVD subjects in each training and validation fold to prevent sampling bias during training and evaluation of the model. The leave-6-subjects-out component guarantees that all samples originating from the same subjects are grouped within a single fold. This prevents information leakage about the subjects into the validation set.
The hyperparameter space is shown in Table B.1 and all possible combinations were explored. The combination with the highest average accuracy using SL6SOCV is used to train a model on the development set, whereafter this model is evaluated on the test set.

Hyperparameter	Values	Description	Naive	M	M+A	M+A+MD	M+A+KL	M+A+C
w_on	300	Search window onset (s)	300	300	300	300	300	300
w_recovery	[600, 900]	Search window recovery (s)	600	600	600	600	600	600
n_hidden	[100, 20, 10]	Neurons in last hidden layer	10	10	10	10	10	10
n_fc	[1,2,3]	Fully-connected layers	2	2	2	2	2	2
n_conv	[0,1,2]	Convolutional layers (conv only)	0	0	0	0	0	0
act_funcs	[Tanh, RELU]	Activation function	Tanh	Tanh	Tanh	Tanh	Tanh	Tanh
b_norm	[True, False]	Batch norm	0	0	0	0	0	0
dropout	[0, 0.6]	Dropout	0	0	0	0	0	0
lr	[0.02, 0.05]	Learning rate	0.05	0.02	0.05	0.05	0.05	0.05
batch size	8192	Batch size	8192	8192	8192	8192	8192	8192
kernel size	[5, 20]	Kernel size (conv only)	N/A	N/A	N/A	N/A	N/A	N/A
pooling size	2	Pooling size, stride 2 (conv only)	N/A	N/A	N/A	N/A	N/A	N/A
channels	16	Channels (conv only)	N/A	N/A	N/A	N/A	N/A	N/A
w_week_size	604800	Aggregation window size (s)	604800	604800	604800	604800	604800	604800
w_week_stride	43200	Aggregation window stride (s)	43200	43200	43200	43200	43200	43200
α	[0, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10]	Divergence loss regularization coefficient	N/A	N/A	N/A	1e-5	1e-5	1e-2
β	[0, 0.1, 1, 10]	Contrastive loss margin coefficient (contrastive loss only)	N/A	N/A	N/A	N/A	N/A	0.1

Table B.1.: Gridsearch hyperparameters explored and the corresponding optimal values for each method. M: mean inactive-peak normalization, A: Weekly Aggregation, MD: Mean Divergence, KL: KL Divergence, C: Contrastive Loss. Naive model uses curve-based classifier only.

The number of neurons in each fully-connected hidden layer linearly decreases based on the input and last hidden layer (n_{hidden}) according to the formula:

$$N_l^n = n_{hidden} + \left\lceil \frac{\left(\frac{w_{on}+w_{recovery}+5}{5} - n_{hidden}\right)}{n_{fc}+1} \right\rceil * l, \quad l = 0, \ldots, n_{fc}+1 \tag{B.1}$$

where N_l^n is the number of neurons in the l-th hidden layer. For example, an input size of 160 neurons, with 10 neurons in the last hidden layer and 2 hidden layers in between would result in a network with 110, 60 and 10 neurons in the first, second and last hidden layer respectively. The fully-connected layers are considered with and without preceding convolutional layers. Each convolutional layer's output is:

$$N_l^k = N_{l-1}^f - w^k + 1 \tag{B.2}$$

where N_l^k is the feature map size after applying the kernel, N_{l-1}^f is the feature map size (or number of neurons) of the previous layer and w^k is the kernel size. Similarly, pooling is applied after applying the kernel with a stride of 2:

$$N_l^f = \frac{N_l^k - w^p}{2} + 1 \tag{B.3}$$

where N_l^f is the feature map size of the l-th layer and w^p is the pooling size.

B

B.2. SENSITIVITY ANALYSIS ON QUANTILE NORMALIZATION

Figure B.1 illustrates quantile normalizations for different quantile values. The minimum (q=0) is perhaps similar within a group but different between the REF and CVD groups. This would make sense as resting heart rate of patients with CVD is characteristically different from healthy subjects. In contrast, higher quantiles that ignore extremeties, do not capture this and are more similar across groups.

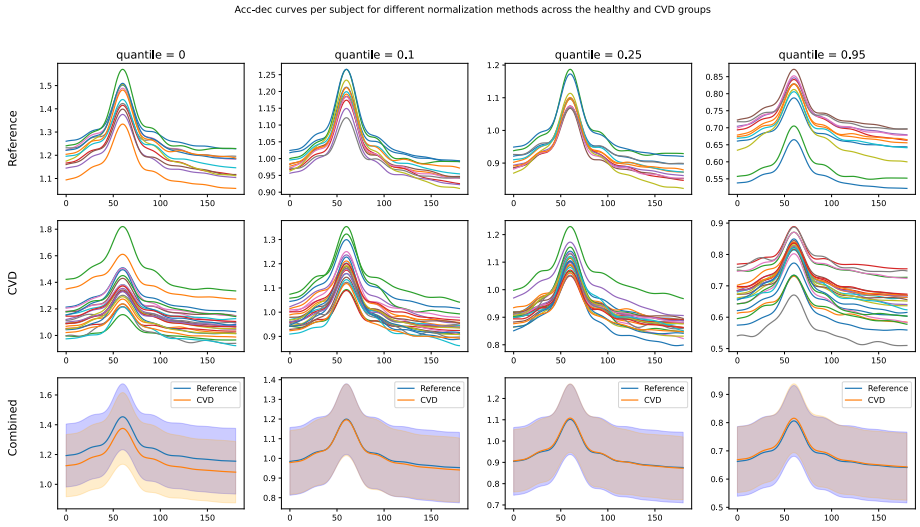


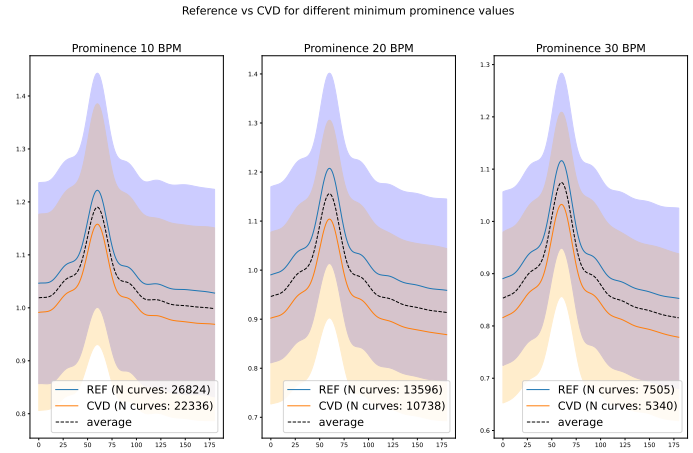
Figure B.1.: Effect of quantile normalization applied per-subject. Every curve in the first two rows represent the average curve from each subject of the Ref and CVD group, respectively. The last row show the average curve per group.

B.3. IMPACT OF PROMINENCE AND ACTIVITY LEVELS

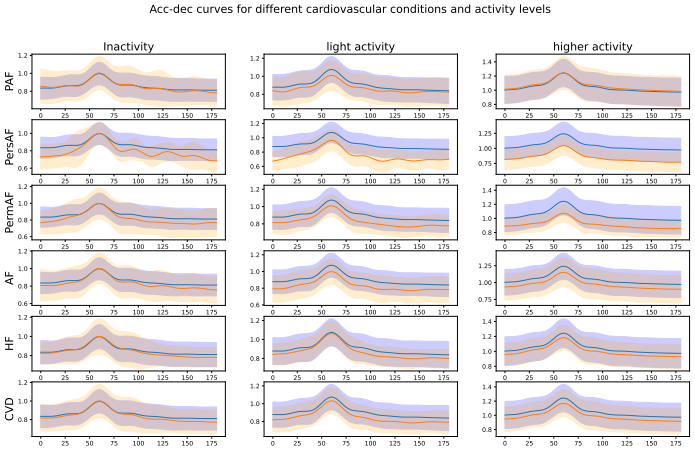
The acc-dec curves are derived from the peak detection algoritm which uses the prominence value to determine whether a peak stands out. To investigate the influence of how much the peak stands out, Figure B.2a shows the average curve for the REF and CVD class for three different different values of the prominence. A prominence level that is either too small (10 BPM) or too high (30 BPM) appears to reduce the distance between the average REF and CVD curve, especially noticeable when comparing either group to the overall mean curve in the region before and after the peak, while 20 BPM is a reasonable compromise.

Next we analysed the influence of activity level on the difference between the CVD and REF group. Figure B.2b shows that, compared to peaks during inactivity (0 steps) and light activity(1-20 steps), difference between acc-dec curves of the CVD and REF group is largest when considering only higher activity (>20 steps). This is true for all CVDs, except for Paroxysmal Atrial Fibrillation (PAF) indicating that it is a more suitable configuration for a ML model.

B



(a)



(b)

Figure B.2.: Effect of a prominence (top, a) and activity level (lower, b) on the acc-dec curves for the REF (blue) and CVD (orange) class. The curves are averaged per class with their corresponding +- 2 standard deviations interval. A:Average REF and CVD curves with a prominence value of 10 (left), 20 (middle) and 30 (right) BPM. Black dashed line denotes average over all curves. B: Effect of different levels of activity. Inactivity, light activity and higher activity correspond to 0, 0 to 20 and higher than 20 recorded steps, respectively. CVDs investigated are PAF (Paroxysmal Atrial Fibrillation), PersAF (Persistent Atrial Fibrillation), PermAF (Permanent Atrial Fibrillation) and HF (Heart Failure). AF (Atrial Fibrillation) contains PAF, PersAF and PermAF. CVD contains all CVDs.

C

APPENDIX OF CHAPTER 5

C.1. BIGIDEASLAB_STEP

In this paper we used the BigIdeasLab_STEP dataset from PhysioNet [3]. This dataset includes data from 53 participants and was recorded in July-August 2019. The age of the participants ranged from 18 to 54. Each person needed to perform three study protocol rounds with different types of wearables. One study protocol round consisted of five activities in the following order:

1. Seated rest (4 min)
2. Paced deep breathing (1 min)
3. Physical activity (5 min)
4. Seated rest (2 min)
5. Typing (1 min)

In the experiment, every person wore all the available devices spread over multiple rounds, capturing different amounts of samples. Round 1:Empatica E4 ($N \approx 140K$), Apple Watch 4($N \approx 13K$). Round 2:Fitbit Charge 2($N \approx 11K$). Round 3: Garmin Vivosmart 3($N \approx 37K$), Xiaomi Miband($N \approx 21K$) and Biovotion Everion($N \approx 161K$). During the whole experiment, the participant always wore a Bittium Faros 180 ECG device($N \approx 221K$) as a reference.

The dataset consists of a synchronised heart rate value in bpm between the smartwatch and the ECG device. Moreover, it is annotated with one of the five activities the person is performing. In the dataset, this is denoted by the labels Rest, Breathe, Activity, Rest after Activity (RestAC) and Type. In the experiments, only the heart rate data is used of the Apple Watch because of its strong correlation with the heart rate time series of the ECG ground truth in comparison with the other wearables.

C.2. CLASSIFICATION MODELS

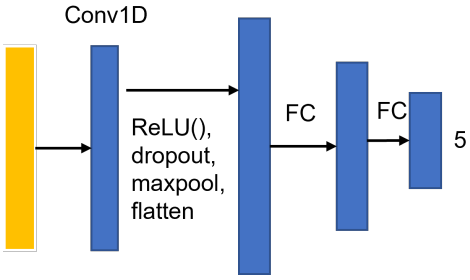
In several experiments, we used a support vector machine (SVM). An SVM tries to maximize the margin between two classes. The SVM maximizes the generalization of a model [4]. For multiclass classification one can use multiple binary SVMs. Two of the methods used for this are One-against-all and one-against-one [5]. For the experiments, we used the implementation provided by scikit-learn, which uses the one-against-one method [6].

In addition to the other experiments, we researched the influence of the addition of handcrafted (HC) features with deep learning models. To investigate this, three different DL models with the addition of HC features were used alongside a DL baseline. All of the DL models started with a 1-D convolution and had three or four fully connected layers.

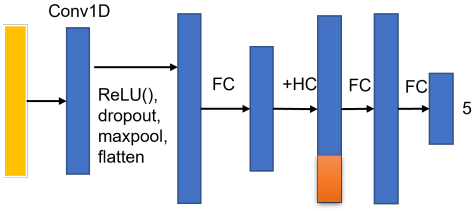
The baseline model starts with a 1-D convolution where the raw sequence input will be processed. Next, it goes through a ReLU, Dropout and Max pooling layer and finally, a flatten layer. After the flattening, it is processed by a fully connected layer, followed by a ReLU and a last fully connected layer to bring the output dimension to the required number of classes. A high-level graphical overview can be seen in Figure C.1a.

The first model is highly similar to the baseline model but it adds an extra layer between the last two fully connected layers. So after the first fully connected layer after flattening, the model adds the HC features to the output of this layer. Next, it processes through another fully connected layer and thereafter it goes through the last fully connected layer. This layer ensures that it ends with the correct dimension. A simple graphical representation can be found in Figure C.1b.

The second model is integrating the HC features directly at the beginning of the DL model. This is achieved by concatenating the HC features with the raw time series input. This results in a larger input vector than with the previous model. The third model is very similar to the first one but with one addition. Instead of adding the HC features directly to the output of the fully connected layer, the HC features first go through a fully connected layer and this output is connected to the output of the first fully connected layer of the model. A graphical representation of both models can be found in Figure C.2.



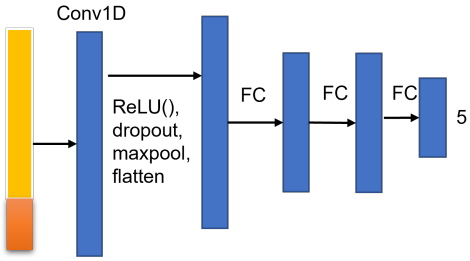
(a) baseline network



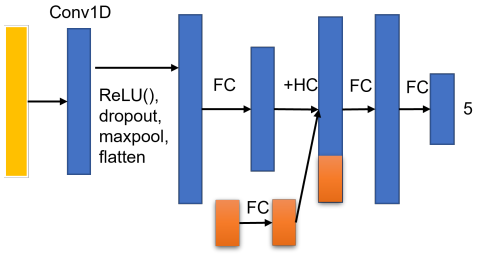
(b) First network

Figure C.1.: High-level overview of the internal working of the baseline DL model(a) and the first DL model (b) that make use of HC features. Yellow represents the raw input sequence and orange represents the HC features.

C



(a) Second network



(b) Third network

Figure C.2.: High-level overview of the internal working of the second(a) and third(b) DL model that makes use of HC features. Yellow represents the raw input sequence and orange represents the HC features.

D

APPENDIX OF CHAPTER 6

D.1. SAMPLING STRATEGIES

Algorithm 1 Stratified Pairwise Sampling

Input: Number of samples per subject N
for each subject S **do**
    Generate class stratified sample vector:
    Create a sample vector $D(S)$ with $\frac{N}{2}$ samples from class 1 and $\frac{N}{2}$ samples from class 2
    Shuffle vector $D(S)$ randomly
end for
Sampling Process:
while number of samples drawn from D is less than N **do**
    for each subject S **do**
        Within-subject sampling:
        Draw pair (x_1, x_2) from sample vector $D(S)$ without replacement for subject S
        Between-subject sampling:
        Draw sample x_3 from sample vector $D(S)$ without replacement for subject S
        Draw sample x_4 from sample vector $D(S')$ with replacement for other subject $S' \neq S$
    end for
end while

Algorithm 2 Stratified Triplet Sampling

Input: Number of samples per subject N
for each subject S **do**
    Generate class stratified sample vector:
    Create a sample vector $D(S)$ with $\frac{N}{2}$ samples from class 1 and $\frac{N}{2}$ samples from class 2
    Shuffle vector $D(S)$ randomly
end for
Sampling Process:
while number of samples drawn from D is less than N **do**
    for each subject S **do**
        Draw anchor and positive item (x_A, x_P) from sample vector $D(S)$ without replacement for subject S
        Draw negative item x_N from sample vector $D(S')$ with replacement for subject $S' \neq S$
    end for
end while

D.2. PAIRWISE SUBJECT FACTORIZATION

To investigate the factorization models' ability to split data in the domain-related and class-related latent space, we consider training the models on two subjects at a time and repeat this training on all pairwise combinations of 10 subjects. After training the model, a logistic regression is used to probe subject separation in the domain latent space and class separation in the class latent space and reporting the train ROCAUC, illustrated in Fig.D.1. The MLP baseline and factorization models are based on a neural network consisting of 10 layers in the encoder, with 200 neurons in the bottleneck layer and 240 neurons in the input layer. The number of neurons in intermediate layers decay linearly. The FAE weighs the class and domain loss equally ($\beta = 0.5$ in Equation 6.3) and does not consider pairs with different domains (i.e. y_{ij}^{d} is always 1 in Equation 6.7) in the domain loss. The GFAE allows for unequal weight distributions for the class and domain loss and also considers a non-zero domain margin. To that end we consider the GFAE both when the domain margin is zero and non-zero. The TFAE additionally uses triplet sampling.

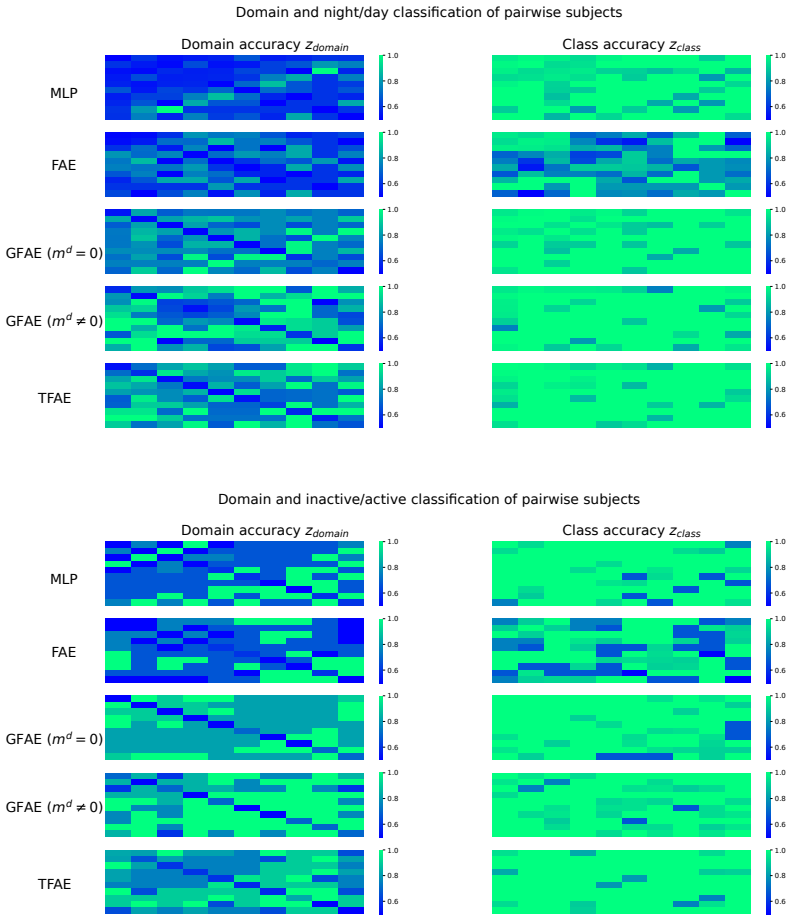


Figure D.1.: Pairwise classification profiles for night/day classification (top) and inactive/active classification (bottom) for (a randomly selected set of) 10 subjects. Every cell represents the train ROCAUC of the subject classification, i.e. the seperability of the pair of subjects in the domain space (left column) and class classification, i.e. the seperability of the classes in the class space (right column) using two subjects. In this way, all pairwise combinations of 10 subjects in total are considered.For the ROCAUC measuring domain separation, we expect the model to learn that subjects are distinguishable within this space. Note that for the subject classification ROCAUC, we expect that the model learns that in this space subjects are separable, so if two different subjects are considered, there should be an AUROC close to 1, and if they are the same (on the diagnoal), the AUROUC is expected to be close to 0.5. On the other hand for the the class classification, we expect that the model learns to separate classes, disregarding the subjects considered.

Overall, Fig D.1 reveals that even when the models achieve a high ROCAUC on the class (right column), for many pairs of subjects the subjects are not classified well (left column). The models do degrade to near random performance on the diagonals, where the same subject is considered in the pair and therefore most similar to each other than any other subject pair. This is to be expected, as finding a latent space where the subjects are seperated, while they are in fact the same subject, is contradicting. But, many off-diagonal pairs (pairs of different subjects) perform poorly, which may indicate that these subjects are similar hindering the model from learning a latent space where they are well seperated. Furthermore, the GFAE performs better than the FAE, with and without the addition of the domain margin m^d. The TFAE performs similar to the GFAE with $m^d = 0$.

The results indicate that the GFAE (both with and without the addition of the domain margin) and TFAE consistently outperform the FAE. This highlights the significance of carefully balancing the class and domain losses to optimize performance. Moreover, explicitly modeling inter-subject differences through the domain margin loss yields the best overall performance profile.

However, while most models can fit the task well, they struggle to learn a domain latent space where subjects are distinctly separable, even within the training data. The corresponding latent spaces learned by the model for one of the pairs is illustrated with UMAP in Figure D.2.

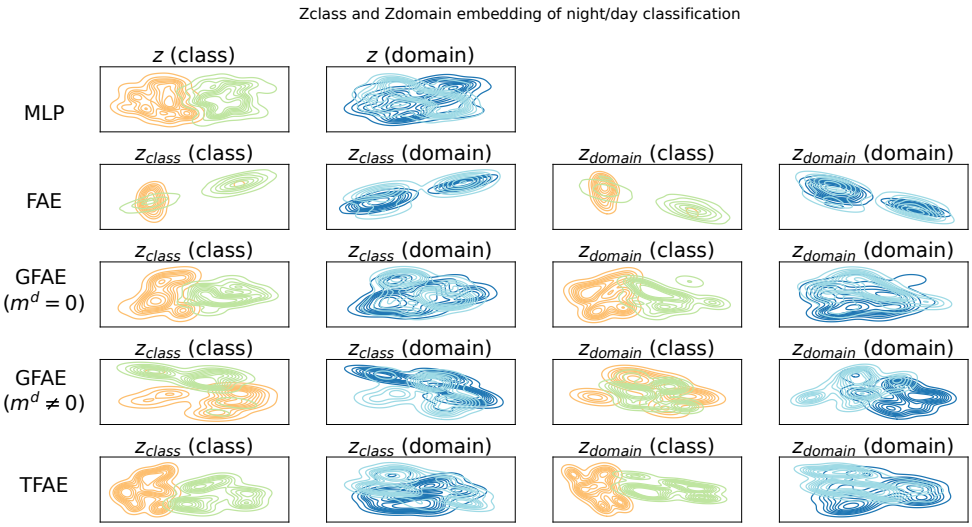


Figure D.2.: UMAP embedding for train set of two subjects.

D.3. MULTI-SUBJECT FACTORIZATION UMAPS

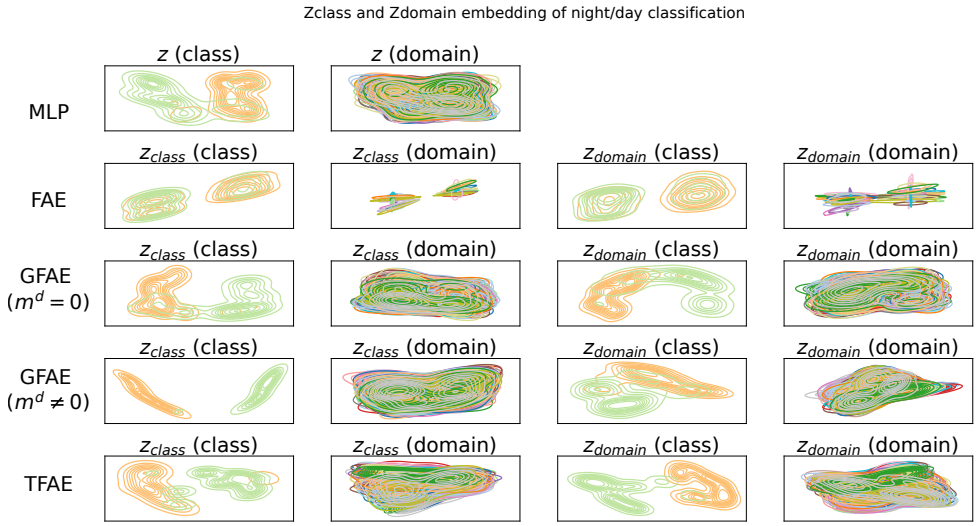
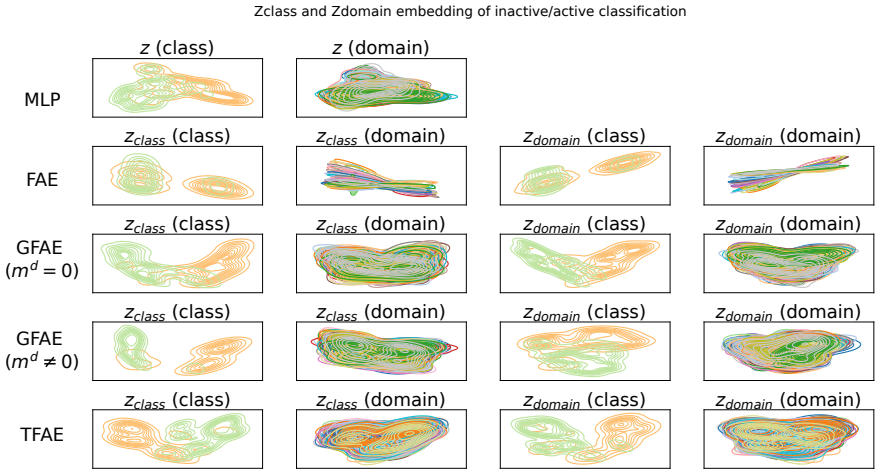
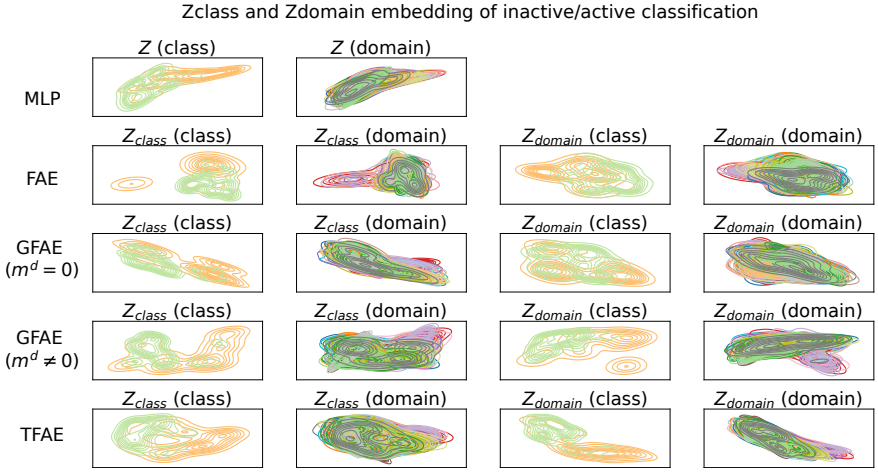


Figure D.3.: Night/day UMAP embedding train set. Parantheses indicate coloring (by class or domain)



(a) Inactive/active UMAP embedding train set. Parantheses indicate coloring (by class or domain)



(b) Inactive/active UMAP embedding test set. Parantheses indicate coloring (by class or domain)

Figure D.4.: UMAP visualization of the models using the training and test data for the inactive/active task.

Table D.1.: Inactive/active class and domain ROCAUC of logistic regression on Z_{class} and Z_{domain}.

Train Data (50 subjects)

Model	Space	Class Acc.	Domain Acc.
MLP	z^c	84.33	7.54
FAE	z^c	85.70	3.71
	z^d	71.31	9.83
GFAE ($m^d = 0$)	z^c	88.43	14.72
	z^d	88.22	17.14
GFAE ($m^d \neq 0$)	z^c	93.34	13.30
	z^d	92.22	33.92
TFAE	z^c	92.22	7.12
	z^d	91.31	23.51

Test Data (30 subjects)

Model	Space	Class Acc.	Domain Acc.
MLP	z^c	84.91	7.80
FAE	z^c	86.02	3.41
	z^d	78.01	7.23
GFAE ($m^d = 0$)	z^c	90.14	7.21
	z^d	86.12	9.84
GFAE ($m^d \neq 0$)	z^c	88.51	12.11
	z^d	88.13	13.22
TFAE	z^c	87.81	12.61
	z^d	84.62	12.93

D.4. HEART ACCELERATION HISTOGRAM

The histogram in Figure D.5 reveals that after applying the (relative) heart acceleration transformation on the entire data set, the timeseries become sparse. For the heart acceleration, almost half of the data points are zero and more than 95% are either zero or differing by 1 BPM. This shows that most of the heart rate dynamics lies in a very small percentage of the data.

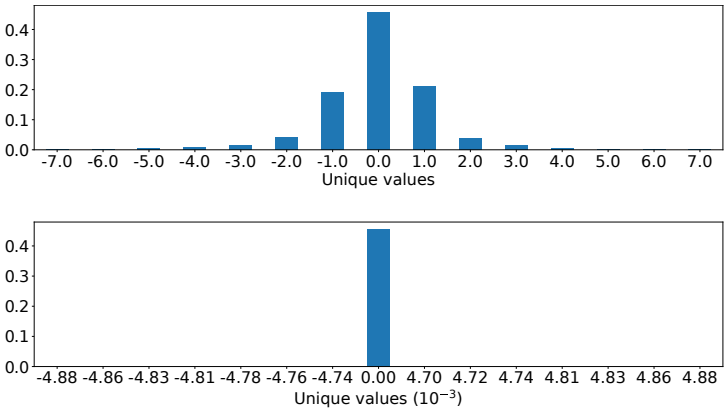


Figure D.5.: Histogram of unique values of heart acceleration (top) and relative heart acceleration (bottom).

D.5. ROC CURVES

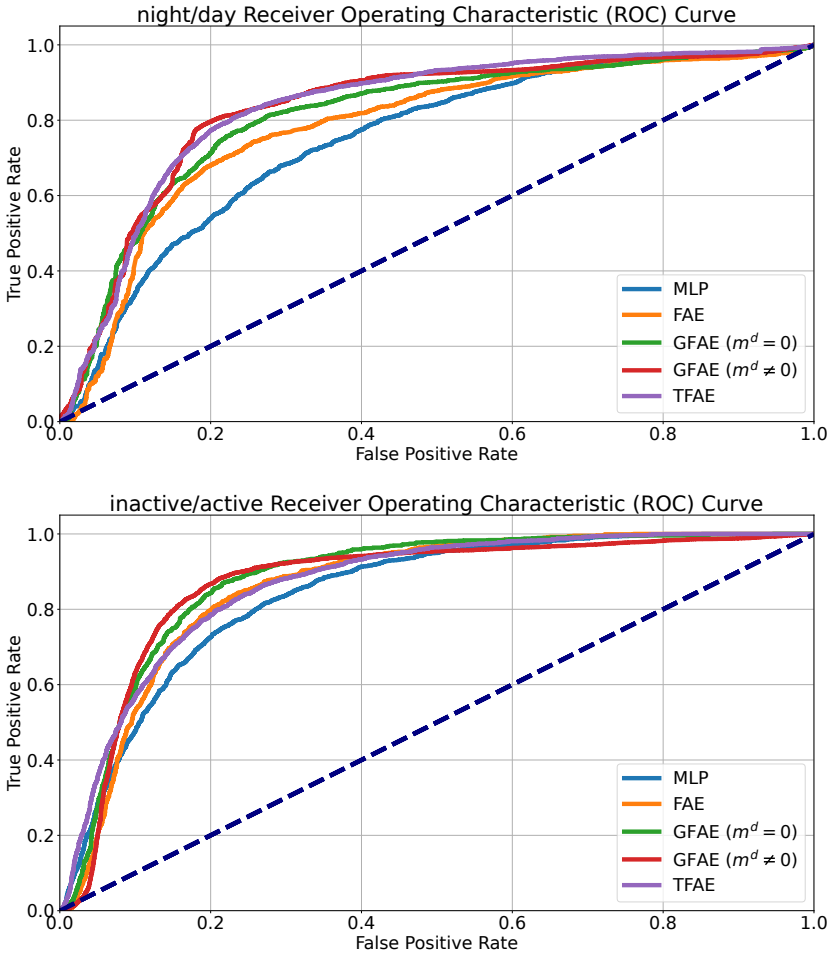


Figure D.6.: ROC curves for night/day (top) and inactive/active (bottom) classification. Dashed line indicates baseline performance (ROCAUC = 0.5).

ACKNOWLEDGEMENTS

I would like to thank **Marcel** for being my promoter. You were essentially one of my supervisors due to your involvement and time investment. Thank you for your critical thinking and guidance. **David**, you have taught me in several courses, were my supervisor during my master's thesis and also my PhD. Thank you for contributing to my growth as a scientist. Your down to earth approach to research and setting up experiments shaped me while also leaving the feeling that there is much more to learn. Last of the supervisors but not least, I would like to thank **Ivo** for his supervision. Your insights and occasional lectures on cardiology were truly essential in shaping our research. Although many delays were present in my PhD, without your assertive getting-things-done attitude, I am sure I would not have been able to finish. Besides contributing in terms of content and process, more importantly, you inspired me with your strong morals, empathy, and leadership, and I aspire to somehow incorporate this in my own 'philosophy of research'. Finishing my PhD leaves a bittersweet taste, as there is so much more that we can achieve together, and therefore I am happy that I can be your 'first mate' a little bit longer. **Mo**, thank you for being my wingman during my PhD. Thanks for listening to all my ranting and complaining (which I know is tiring) and also for the valuable discussions we had. Every discussion/course with you taught me something new and never felt like a heated debate. **Xiangwei**, thanks for letting me tease you all the time. From the start of your PhD, you have been very successful, but have always remained playful and modest. Thank you for inspiring me with this. **Yuko-san**, thank you for contributing to an amazing atmosphere with fun discussions in our office. You truly turned it into my comfort zone. **Jin**, I'm not even sure if you are still alive, but if you somehow read this, thank you too for being an amazing office mate and inspiring me with our varied discussions from language to statistical pattern recognition. A big shoutout to the Diamond Dogs **Taylan**, **Tiffany** and **Ojas**. Our discussions (both at the office and walking to the train station) meant a lot to me. Ojas and Tiffany, I would be your number 1 fan if you guys started a podcast. Taylan, I appreciate what Ojas accurately describes as *shiftiness* and thank you for inspiring me with your rebellious opinions and amazing engineering skills. My dream is to learn Rust together with you. **Chirag**, thank you for all your fatherly advice. Our bike trip together with **Yeshwanth** was a bittersweet experience that I cherish. Yeshwanth, thank you for your kindness and innocence. Laughing at bad

jokes together with you at Haga, made it much more bearable to finish the final pieces of my PhD. **Osman**, thank you for your varied advice, from music streaming, to relationships, to science. Your energy is always infectious, and I hope to be able to enjoy it during our work together as long as possible. **Ramin**, thank you for always inspiring me with your strong values, caring attitude, and your amazing food recommendations. **Matthijs**, although I didn't have the pleasure (yet ;)) to do scientific research with you, thank you for being a f(un) co-worker to work with and for talking about our strong common desire to learn new things. **Jan**, your managing style and philosophy resonates deeply with me and was a huge factor in smoothly transitioning from my PhD to working in your team. Thank you. **Stavros** and **Tom**, thank you guys for supervising students together with me at the start of my PhD. I learned a lot from you in the process and I considered you my mentors. **Ziqi**, thanks for having breathers with me and Tom. More importantly, I value the serious, yet natural, discussions we had and your raw, honest opinion on matters. I secretly hope, you will return to the Netherlands one day! **Stephanie**, thanks for being my amateur therapist and for letting me vent about things. Watching Boondocks and movies together with you at the office was also fun. **Yasin**, thanks for hanging out with me at the office and going on a mountain hiking trip. I think Mo and I learned everything about Turkey thanks to you. I hope one day we can go hiking there together. **Mahdi**, **Ramin** and **Hesam**, thank you guys for making my Farsi a little bit less bad. Mahdi, thank you for broadening my (and the entire group's) horizon in terms of spirituality, poetry and science. Hesam, thank you for letting me barge into your office once in a while to have a stress-relieving discussion with you. Your calm and kindness helped me push through. Also it was nice to have someone to enjoy eating meat with.

CURRICULUM VITæ

Arman Naseri Jahfari

EDUCATION

2009–2016	Bachelor of Science in Electrical Engineering Delft University of Technology
2016–2019	Master of Science in Electrical Engineering Specialization: Signals & Systems Delft University of Technology
2025	PhD. Computer Science Delft University of Technology *Thesis:* Improving Remote Cardiovascular Care with Wearable Data: Algorithms, Study Design, and Subject-Specific Adaptation *Promotor:* Prof. dr. ir. M.J.T. Reinders

LIST OF PUBLICATIONS

6. K. G. Schmahl, T. J. Viering, S. Makrodimitris, A. Naseri, D. Tax, and M. Loog. “Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings”. In: *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*. 2020, pp. 94–103

5. A. Naseri, D. Tax, M. Reinders, and I. van der Bilt. “Machine learning for cardiovascular outcomes from wearable data: systematic review from a technology readiness level point of view”. In: *JMIR medical informatics* 10.1 (2022), e29434

4. A. Naseri, D. Tax, P. van der Harst, M. Reinders, and I. van der Bilt. “Data-efficient machine learning methods in the ME-TIME study: Rationale and design of a longitudinal study to detect atrial fibrillation and heart failure from wearables”. In: *Cardiovascular Digital Health Journal* 4.6 (2023), pp. 165–172

3. M. Beekhuizen, A. Naseri, D. Tax, I. van der Bilt, and M. Reinders. “Improving performance of heart rate time series classification by grouping subjects”. In: *arXiv preprint arXiv:2311.13285* (2023)

2. A. Naseri, D. M. Tax, M. Reinders, and I. van der Bilt. “Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data”. In: *Heliyon* 10.21 (2024)

1. A. Naseri, D. Tax, I. van der Bilt, and M. Reinders. “Tackling inter-subject variability in smartwatch data using factorization models”. In: *Submitted to Scientific reports* (in review)