Designing and Evaluating Mobile Health Technology for Ambulatory Monitoring and Diagnosis of Heart Arrhythmias

Kumar, Devender

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
Doctor of Philosophy

DTU Health Tech
Department of Health Technology

Designing and Evaluating Mobile Health Technology for Ambulatory Monitoring and Diagnosis of Heart Arrhythmias

Devender Kumar

Kongens Lyngby 2021

DTU Health Tech
Department of Health Technology
Technical University of Denmark

Ørsteds Plads
Building 345C
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
healthtech-info@dtu.dk
www.healthtech.dtu.dk

Summary

Cardiac arrhythmias comprise a large class of cardiovascular diseases (CVD). They are associated with an increased risk of heart attack and stroke and are a leading cause of global deaths. Atrial fibrillation (AF), a type of arrhythmia, alone affects over 1% of the worldwide population. It costs €73–95 million in the annual healthcare budget of Denmark and $26 billion in the USA. Electrocardiogram (ECG) analysis is a cost-effective and non-invasive means for detecting heart arrhythmias. However, due to their sporadic or paroxysmal nature, they often remain undetected in routine in-hospital ECG examinations, and therefore require longitudinal monitoring in patients under free-living conditions. Although longitudinal arrhythmia screening under free-living conditions can help in early diagnosis, it faces several challenges such as sustained patient engagement, poor signal quality, recall bias in the patient-reported symptoms diary or events, and a high false positive rate (FPR) in computer-aided automatic arrhythmia detection algorithms. Furthermore, in free-living ambulatory monitoring, motion artifacts often mimic arrhythmias and cause misdiagnosis. Without an understanding of the patient's ambulatory context, it is difficult to ascertain if the ECG morphology is due to artifacts or arrhythmias. The high FPR in longitudinal screening increases the workload of clinicians (as it requires manual review) and could also lead to over-diagnosis and patient anxiety.

This dissertation investigates the role of context-awareness obtained via mobile and wearable devices to improve ambulatory arrhythmia analysis and reduce the FPR under free-living conditions. In addition, we also address the issues of recall bias in the patient-reported symptoms diary and events, and sustained patient engagement in longitudinal arrhythmia screening.

First, we identified the context information relevant for improving ambulatory arrhythmia screening under free-living conditions. After that, *mCardia* – a context-aware ECG collection system – was designed for longitudinal arrhythmia screening. We evaluated its usability and clinical feasibility in collecting contextualized ECGs for longitudinal arrhythmia screening in patients under free-living ambulatory conditions. Two clinical case studies from the collected contextualized data demonstrated the usefulness of contextual data in improving the manual analysis of ECG.

Furthermore, to improve the automated arrhythmia detection algorithm, we first investigated the influence of the patient's ambulatory context on FPR in a state-of-the-art arrhythmia detection algorithm. The investigation revealed that three specific ambulatory contexts, namely change in body position, activity change, and sudden

movement acceleration, caused a significant number (62%) of non-trivial small segments of false positives. Based on these findings, we proposed a hybrid arrhythmia (AF) detection model named *DeepAware*. It employs deep learning and context-aware heuristics that significantly reduce the FPR under free-living conditions. When used in a clinical setting, *DeepAware* can significantly reduce cardiologists' workload of manual review of FPs, allowing them to focus more on treatment than diagnostics.

With the two clinical case studies and performance of the *DeepAware* model, this PhD thesis demonstrated that contextual information could help in improving both manual and automated arrhythmia detection under free-living ambulatory conditions.

In addition, this dissertation also contributed a 259-day-long contextualized single-channel ECG arrhythmia dataset from patients under free-living ambulatory conditions. This database will help the broader deep learning community in building and evaluating the arrhythmia detection models that can realistically work under free-living conditions. Furthermore, it will pave the way for making the deep learning-based, end-to-end arrhythmia detection models more explainable and will help identify the source of algorithm errors that otherwise remain a black box.

Resumé (Danish)

Hjertearytmier omfatter en række forskellige typer af hjerte-kar-sygdomme. De er forbundet med en øget risiko for hjerteanfald og slagtilfælde og er den primære årsag til dødsfald globalt. Atrieflimren (AF), som er en form for arytmi, påvirker alene over 1% af verdens befolkning. Behandling af AF koster årligt i Danmark mere end 500 millioner kroner og mere end 26 milliarder dollars i USA. Analyse af elektrokardiogram (EKG) er et omkostningseffektivt og ikke-invasiv metode til påvisning af hjertearytmier. Men på grund af arytmiernes sporadiske eller paroxysmale natur, forbliver de ofte uopdagede i rutinemæssige EKG-undersøgelser på et hospital. For at opdage mulige hjertearytmier kræver det derfor ofte, at man overvåger patienten ambulant og over en længere periode, oftest i deres dagligdag. Selvom ambulant screening af arytmier kan hjælpe med tidlig diagnose, står teknologien overfor en række udfordringer, herunder dårlig signalkvalitet, manglende eller forkert rapporteringer af symptomer fra patienten, og en høj andel af falsk-positive ved computerstøttede algoritmer til automatisk arytmi-identifikation. Desuden ligner bevægelsesartefakter under naturlige omgivelser arytmier og forårsager ofte en fejldiagnose. Uden forståelse for patientens kontekst er det vanskeligt at fastslå, om EKG-morfologien skyldes artefakter eller arytmier. Den høje andel af falsk positive øger klinikernes arbejdsbyrde (da disse kræver manuel gennemgang), og kan føre til overdiagnosticering og patientangst.

Denne afhandling undersøger om indhentning af viden om patientens hverdagskontekst via. mobile og bærbare enheder sammen med EKG mållinger har et potentiale for at forbedre ambulant arytmi-analyse og reducere antallet af falsk positive. Afhandlingen behandler også udfrodringerne omkring patientens erindring og rapportering af symptomer samt hvordan man kan sikre et vedvarende engagement af patienten i længerevarende arytmi screeninger.

Afhandlingen præsenterer fire hovedresultater. For det første identificerede den en række kontektuelle parametre, som er relevante at indsamle for at forbedre ambulant arytmi-screening. For det andet præsenterer den *mCardia*, som er en smartphone-baseret teknologi til arytmi-screening til brug i patientens dagligdag. Gennem en række patientstudier, demonstreres teknologiens brugermæssige og kliniske anvendelighed til indsamling af kontekstualiserede EKG data for ambulant arytmi-screening. I to kliniske casestudier demonstreres nytten af at bruge kontekstuelle data til forbedring af den manuelle analyse af EKG. For det tredje præsenteres en analyse, der viser at hvis man anvender nuværende "state-of-art" maskinlæringsalgoritmer til automatisk identifikation af AF på det indsamlede ambulante EKG data, resulterer dette i en

meget høj rate (62%) af falske positive. Afhandlingen præsenterer en undersøgelse af årsagen til dette og finder at dette ofte skyldes tre specifikke ambulante sammenhænge; nemlig; ændring i kropsposition, aktivitetsændring og pludselig bevægelsesacceleration. For det fjerde præsenterer afhandlingen en maskinlæringsmodel for automatisk identifikation af AF som bygger på disse fund. Denne model har fået navnet *DeepAware*. Modellen anvender "deep learning" metoder kombineret med heuristikker der anvender viden om patientens dagligdagskontekst. En nærmere analyse viser, at *DeepAware* er i stand til at reducere antallet af falsk positive betydeligt.

Med udviklingen af *mCardia* og *DeepAware* modellen kombineret med de patientrettede studier, de kliniske casestudier, samt analysen af det indsamlede data har denne afhandling vist, at kontekstuelle oplysninger kan hjælpe med at forbedre både manuel og automatiseret arytmieidentifikation under ambulante forhold. Derudover har afhandlingen bidraget med et arytmi-datasæt, som rummer 259 dages kontektuel EKG data optaget ambulant fra 24 patienter. Datasættet vil kunne hjælpe forskere og studerende med at opbygge og evaluere modeller for identifikation af arytmier, der realistisk afspejler optagelser fra patientens dagligdag. Desuden vil det kunne bane vejen for at gøre deep learning-baserede arytmi-identifikationsmodeller mere gennemsigtige og vil hjælpe med at identificere kilder til algoritmefejl, der ellers normalt forbliver en 'sort boks'.

Preface

This PhD dissertation was prepared at the Department of Health Technology and Copenhagen Center for Health Technology (CACHET) at the Technical University of Denmark to fulfill the requirements for acquiring a PhD degree.

The PhD project was a part of the REAFEL project for optimizing diagnosis of atrial fibrillation in frail elderly patients, funded by the Innovation Foundation of Denmark under grant 6153-00009B. The thesis was completed under the supervision of Professor Jakob E. Bardram and Professor Sadasivan Puthusserypady from DTU Health Tech. It includes one published and one accepted journal article, two under review journal articles, and one journal article manuscript in preparation.

Kongens Lyngby, May 14, 2021

Devender Kumar

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Jakob Eyvind Bardram and co-supervisor Professor Sadasivan Puthusserypady for their invaluable advice and continuous support during the journey of my PhD. Their extensive expertise and plentiful experience have supported me in my academic research and daily life (especially after being disrupted by COVID-19). I would also like to thank cardiologist Dr. Helena Dominguez and the entire team at the Holter Laboratory in Bispebjerg and Frederiksberg Hospital for their support during the design and evaluation of *mCardia*. Dr. Dominguez's help immensely improved my understanding of the clinical side of the arrhythmia monitoring process. I am thankful for our long brainstorming sessions, workshops, and design meetings.

I am extremely grateful to Dr. G.N. Saxena, Prof. Bipinkumar Rathod, and Dr. A. K. Singh from Mahatma Gandhi University of Medical Sciences & Technology, Jaipur, India, for providing the necessary support and approval for *mCardia*'s feasibility study in India. Also, special thanks to Dr. Rajeev Sharma and his team at the Department of Cardiology for helping in patient recruitment. I would also like to thank Dr. Kamal Sharma, a senior interventional cardiologist at U.N. Mehta Institute Of Cardiology, for logistical support during *mCardia*'s feasibility study in India and technical support in ECG annotation and validation of my algorithms.

I want to thank the entire CACHET team, fellow PhD students Raju Maharjan, Pegah Hafiz, Giovanna Nunes Vilaza, and postdoc researchers Steven Jeuris, Darius Adam Rohani, and Alban Maxhuni for making my PhD journey a wonderful and memorable experience. Special thanks go to student assistants and programmers Andrea Quemada Lopategui, Malte Moser Kampmark, and Barnabás Homola for their assistance in the development of *mCardia*.

I would like to express my gratitude to my parents and HH RNS – my friend, philosopher, and guide. Without your tremendous encouragement in the past few years, it would not have been possible to complete my study.

Finally, I am grateful to Innovation Fund Denmark and CACHET for funding this research.

Contents

Summary i

Resumé (Danish) iii

Preface v

Acknowledgements vii

Contents ix

Acronyms xiii

List of Figures xvii

List of Tables xxi

1 Introduction 1
- 1.1 Context and motivation . . . 1
- 1.2 Problem statement . . . 3
- 1.3 Research methods . . . 5
- 1.4 Research contribution . . . 9
- 1.5 Scientific publications in thesis . . . 10
- 1.6 Thesis overview . . . 12

2 Background and Related Work 13
- 2.1 Medical background . . . 13
  - 2.1.1 Electricity of the heart . . . 14
  - 2.1.2 What are arrhythmias and how are they diagnosed? . . . 14
  - 2.1.3 The ECG and its different components . . . 14
  - 2.1.4 Atrial fibrillation and ECG morphology during atrial fibrillation 15
- 2.2 ECG monitoring for arrhythmia detection . . . 16
  - 2.2.1 ECG monitoring under ambulatory free-living conditions . . . 17
  - 2.2.2 Other modality of ambulatory arrhythmia monitoring . . . 17
  - 2.2.3 Challenges in longitudinal arrhythmia screening under free-living conditions . . . 18

2.3 Systems for collecting contextualised mobile ECG 19
2.4 ECG database for evaluation of arrhythmia detection algorithms . . . 21
2.4.1 Limitation of existing public arrhythmia dataset: 24
2.5 Computer aided arrhythmia diagnosis 24
2.5.1 Algorithms based on feature engineering 25
2.5.2 Algorithms based on end-to-end deep learning 26
2.5.3 Overview of deep learning architectures used in ECG classifications 27
2.5.3.1 State-of-the-art deep learning-based algorithms for arrhythmia detection 30
2.6 Research gaps 35

3 Technology for Collecting Contextualized ECG for Arrhythmia Screening under Free-Living Conditions 37
3.1 Motivation 38
3.2 Research methods 38
3.2.1 Identifying the context of use 39
3.2.2 Requirements specification 40
3.2.3 Iterative system design and mCardia user interface 41
3.3 Architecture and implementation 43
3.4 Summary 44

4 Usability and Clinical Feasibility Study of *mCardia* 47
4.1 Study design 47
4.1.1 Patients recruitment 48
4.1.2 Study procedure 48
4.2 Results 48
4.2.1 User experience 49
4.2.2 Performance of data collection 52
4.2.3 Qualitative analysis of post-study interview 53
4.2.4 Clinical usefulness of the contextual data 55
4.3 Summary 55

5 Impact of Ambulatory Contexts on False-Positive Rate: An Investigation 57
5.1 Motivation 57
5.2 Research method 58
5.3 Results 61
5.4 Summary 65

6 CACHET-CADB: Contextualized ECG Database for Arrhythmia Screening 67
6.1 Motivation 67
6.2 Design, development, and validation 69
6.2.1 ECG annotation tool 70
6.3 Contributions and summary 70

7 DeepAware: Deep Learning and Context-Aware Heuristic base Hybrid Model for AF Detection **73**
7.1 Motivation . . . 73
7.2 Architecture and implementation . . . 74
7.3 Datasets . . . 77
7.4 Results . . . 77
7.5 Summary . . . 80

8 Discussion **81**
8.1 Relevant contexts and designing tool for longitudinal arrhythmia screening . . . 82
8.2 Impact of ambulatory contexts on FPR . . . 83
8.3 Contextual information for improving automatic arrhythmia detection algorithm . . . 84
8.4 CACHET-CADB for explainable deep learning models . . . 85
8.5 Challenges in context-aware arrhythmia monitoring . . . 86
8.6 Limitations . . . 86
8.7 Future work . . . 87

9 Conclusion **89**

A Journal Papers **91**
A.1 Mobile and Wearable Sensing Frameworks for mHealth Studies and Applications: A Systematic Review . . . 91
A.2 mCardia: An ambulatory Context-Aware ECG Collection System for Arrhythmia Screening . . . 120
A.3 Contextual and Temporal Distribution of False Positives in a Deep Learning-Based Atrial Fibrillation Detection Algorithm: An Investigation 149
A.4 CACHET-CADB: A Contextualized Ambulatory Electrocardiography Arrhythmia Dataset . . . 172
A.5 DeepAware: A Hybrid Deep Learning and Context-Aware Heuristics Based Model for Atrial Fibrillation Detection . . . 188

B List of CUMACF Questionnaires in mCardia's usability and feasibility study **199**

Bibliography **201**

Acronyms

Acc accuracy. 7–9, 73, 77

AF atrial fibrillation. xviii, xix, xxi, 1, 3, 7–12, 14, 21, 25, 31, 33, 41, 56–61, 63–66, 70, 71, 73, 74, 77, 80, 82–86, 89, 90

AFDB MIT-BIH AF Database. xix, 21, 22, 31, 33, 34, 59, 61, 67, 74, 77, 79, 84

AFL Atrial Flutter. 31, 33

API application programmer interface. 44

BLSTM Bidirectional Long-Short Term Memory. 31, 33

BTLE Bluetooth low energy. 44

CACHET-CADB CACHET Contextualised Arrhythmia Database. xix, xxi, 8, 10–12, 36, 60, 63, 67–71, 77, 79, 80, 82, 85, 87, 90

CACHET-NSRDB CACHET Contextualised Normal Sinus Rhythm Database. xxi, 77, 80, 82, 85, 90

CAH context-aware heuristics. xix, 74–76

CAMS CARP Mobile Sensing. xviii, 6, 44, 46

CinCDB Computing in Cardiology Challenge 2017 Dataset. 22, 31, 33, 34, 67

CM confusion matrix. 8, 9

CNN Convolutional Neural Network. xvii, 26, 27, 31–34, 60, 76, 78

CUDB CU Ventricular Tachyarrhythmia Database. 22, 23, 33, 34

CUMACF CACHET Unified Methodology for Assessment of Clinical Feasibility. xviii, 6, 49

CVD cardiovascular diseases. i, 1

DBN Deep Belief Network. 26

DeepQ DeepQ Arrhythmia Database. 22

DL deep learning. xviii, xix, xxi, 2, 7–10, 12, 13, 26–28, 30, 31, 35, 36, 56–61, 63–67, 69, 71, 73, 74, 80, 82, 84–87, 89, 90

ECG electrocardiogram. i, xviii, xix, 1–4, 6–12, 14, 17–20, 35, 38, 40, 41, 45, 51, 53, 54, 57, 58, 61, 65, 74–76, 81–84

EMA ecologically momentary assessment. 41

ESTDB European ST-T Database. 31, 34

FE feature engineering. 2, 24, 25, 57

FP false positive. ii, xix, 4, 5, 7–10, 35, 58–61, 63–66, 68, 71, 73, 74, 76, 82, 84, 85, 89, 90

FPR false positive rate. i, ii, xi, xviii, 2–5, 7–12, 22, 31, 35, 36, 56–59, 61, 63–66, 71, 73, 74, 77, 80–85, 89, 90

GRU Gated Recurrent Units. 26, 28, 29

HR heart rate. xvii, xviii, 3, 42, 45, 54, 55

HRV heart rate variability. xvii, 42, 45

LSTM Long-Short Term Memory. xvii, xviii, 26, 28, 29, 31, 33, 34, 60, 76

LTAFDB Long Term AF Database. 33

MET Metabolic Equivalent for Task. xvii, 42

mHealth mobile health. 6, 9, 71

MITDB MIT-BIH Arrhythmia Database. xix, xxi, 21, 22, 31–34, 59, 61, 67, 74, 77, 79

ML Machine Learning. 10

MLP Multilayer Perceptron. 26

MVFDB MIT-BIH Malignant Ventricular Arrhythmia Database. 22, 23, 31, 33, 34

NSR normal sinus rhythm. xxi, 1, 7, 10, 11, 22, 31, 33, 70, 71, 77, 86

NSRDB MIT-BIH NSR Database. xxi, 22, 31, 33, 34, 59, 61, 77, 79

NSTDB MIT-BIH Noise Stress Test Database. 22, 23, 33

OA-ADB Open-Access Arrhythmia Database. 21, 67

PCD patient-clinician-designer. 5, 38, 41

PPG photoplethysmography. 17, 18

PVC premature ventricular contractions. 74

QTDB QT database. 76, 87

RNN Recurrent Neural Network. xvii, 26, 28, 31, 34, 60, 76

RRI RR-interval. xviii, 31, 32, 60, 73, 74, 76, 77, 80, 84–86

Se sensitivity. 7–9, 77

Sp specificity. 7–9, 73, 77

SVM support vector machine. 31

SVT Supraventricular Tachycardia. xviii, 54, 55, 81

TP true positive. 7, 8, 60, 61, 65

UCD user-centered design. xvii, 5, 38, 40

UI user interface. xvii, 42

UX user experience. 4, 12

List of Figures

1.1 An outline of the dissertation. The color-coding separates the activities performed at each level: Theory, Design, and Observation. Each activity in the rectangle is annotated with the article number in which it is described. UCD: User-centered design, LRW: Literature review. 6
1.2 Major tasks performed to design and develop CACHET-CADB 8

2.1 Cardiac conduction system of the heart 13
2.2 Various components of an ECG signal 15
2.3 ECG morphology during the AF: (1) missing P-waves, (2) and irregular RR intervals. 16
2.4 Ambulatory ECG systems based on monitoring scheme 17
2.5 The workflow of the outpatient arrhythmia screening process (image from [A.2]). 20
2.6 Traditional feature engineering-based approach of arrhythmia detection. . 25
2.7 Deep learning-based end-to-end arrhythmia classification. 26
2.8 (a) Most frequent deep learning architectures used for arrhythmia classification in literature (image data source [43]). 27
2.9 Architecture of the CNN. 27
2.10 Architecture of the RNN. 28
2.11 Structure of the bidirectional LSTM. 29
2.12 Deep Belief Network. 29
2.13 Network structure of Multilayer Perceptron with N inputs, C outputs and L Hidden layers. 30
2.14 Gated Recurrent Unit. r: reset gate, z: update gates, h: activation $\widetilde{h}$: candidate activation. 30

3.1 Schematic diagram of *mCardia* system. The contextual and physiological data collected from ECG Holter and phone are transmitted from the phone to a cloud-based data management system (image from [A.2]). 39
3.2 Timeline of the UCD process. SR: Specify Requirements (image from [A.2]). 40
3.3 Home screen of *mCardia* (final design). The 'wheel' is the main UI element which presents the detailed recordings of HR, HRV, and MET-level in a 24-hour clock (images from [A.2]). 42

3.4 The user interface design of the *mCardia* mobile app. (a) Event details, (b) List of events filled or created by the patient, (c) Daily Info (image from [A.2]). 43
3.5 The architecture of the mCardia system. The green components are part of CARP Mobile Sensing (CAMS) framework, and the blue components are its associated Flutter plugins. Movisens is a single-channel ECG Holter used in the current prototype of the mCardia system (image from [A.2]). . 46

4.1 An outline of the *mCardia*'s feasibility study procedure. Each pink box denotes a single day in the study duration. CUMACF: CACHET Unified Methodology for Assessment of Clinical Feasibility [16, 15]. The CUMACF questionnaires are available in Appendix B. 49
4.2 Patients using *mCardia* system in their free-living conditions. 50
4.3 The CUMACF [16] based perceived usability and usefulness scores of *mCardia* system. The questionnaires (Q1–Q19) are provided in the Appendix B (image from [A.2]). 51
4.4 Per participant number of annotated or deleted/unfilled events (image data from [A.2]). 52
4.5 Data collection yield of various physiological and contextual data (image data from [A.2]). 53
4.6 Clinical case #1 demonstrating the usefulness of context information in ECG interpretation. Although both ECG snippets from the same participant have HR>130, the context information (sleeping or running) helps in distinguishing that (b) is a case of Supraventricular Tachycardia (SVT), whereas (a) is normal (image from [A.2]). 54
4.7 Clinical case #2 demonstrating the usefulness of contextual information in medication prescription. (a) Subject tapped on device and registered an event. (b) Subject gets up from the bed and starts walking (image from [A.2]). 56

5.1 Experimental workflow for investigating the influence of ambulatory contexts on FPR. CACHET-CADB is the contextualized ECG dataset collected during the mCardia's clinical feasibility study (image from [A.3]). . 59
5.2 The network consisted of 2 convolutional layers, followed by a pooling layer. Features extracted in the convolution are fed into the LSTM layer, which consists of 100 hidden units. Finally, the sigmoid layer gives the binary output of the entire 30 RRIs long window. The network takes 30 RRIs (calculated from raw ECG) data as input and outputs a binary classification of AF or non-AF class (image from [A.3]). 60
5.3 Mapping DL-based AF detection algorithm's output to raw ECG timestamp for 24 hours (image from [A.3]). 61
5.4 Annotations summary after manual examination (image from [A.3]). . . . 63

5.5 User context and false positive occurrences in a 24 hour ECG: (a) shows the AF detected by the DL-algorithm in 24 hours ECG, and (b) shows the ground truth of AF episodes after manual annotation. The short segments ($\leq$ 50-seconds) of FP detected by the DL algorithm in (a) are associated with movement accelerations peaks in Figure (c) and the body position and activity change in Figure (d) and Figure (e) (image from [A.3]). 64

6.1 ECG annotation tool (image from [A.4]). 69
6.2 Schematic diagram of ECG annotation platform 70

7.1 The architecture of the proposed *DeepAware* model consists of the following six sub-components: (1) ECG data preprocessing, (2) segmentation, (3) the RR-Net, which take inputs of the RR interval series, (4) the *DENS-ECG*, which takes the raw ECG inputs and gives P-wave count, (5) and a CAH model, which takes user context in a case of ambulatory ECG for dynamically fine-tuning the final output, and (6) Decision box for final binary output (image from [A.5]). 75
7.2 Confusion Matrix on AFDB (image from [A.5]). 79
7.3 Confusion Matrix on MITDB (image from [A.5]). 79
7.4 Confusion Matrix on CACHET-CADB (image from [A.5]). 79

List of Tables

2.1 Specifications and ECG annotations statistics of publicly available arrhythmia databases. Ch: Number of ECG channels, Freq (Hz): sampling frequency. 23
2.2 R-class: No of rhythm/arrhythmias classified, STFT: Short-term Fourier transform, SWT: Stationary wavelet transform, AFL: Atrial flutter, SR: sinus rhythm, DWF: Derived wavelet frames, LS: Logarithmic spectrogram, IHR: Instant heart rate sequence. 33

4.1 An overview of the inclusion and exclusion criteria for the study. 48
4.2 Demographics of the patients (Table from [A.2]). 49
4.3 Issues identified and their relative distribution (Table from [A.2]). 54

5.1 Performance of the used algorithm on public datasets and its comparison with other state-of-the-art algorithms. Ch: number of ECG channels (Table from [A.3]). 62
5.2 Statistics of DL algorithm's performance on CACHET-CADB. AF-DL: number of segments detected as AF by DL algorithm (Table from [A.3]). . . . 65

6.1 ECG annotation statistics in CACHET-CADB. Rhythm class code (1: AF, 2: NSR, 3: Noise, 4: Any other rhythm) (Table from [A.4]) 70

7.1 Performance of the *DeepAware* model on public datasets and its comparison with other state-of-the-art models. Ch: Number of ECG channel (Table from [A.5]). 78
7.2 Performance comparison between RR-Net and *DeepAware* on MITDB and NSRDB datasets (Table from [A.5]). 79
7.3 Performance of *DeepAware* on CACHET-CADB (Table from [A.5]). . . . 80
7.4 Performance of *DeepAware* on CACHET-NSRDB. Each record comprising of 24 hours long contextualised ECG under free living ambulatory conditions. Inputs No: Number of (30x1) input windows. (Table from [A.5]). . 80

CHAPTER 1

Introduction

"When you find someone that makes your heart skip a beat, stop the search and take the risk."— unknown. Beware! it could be arrhythmia in disguise.

• • •

1.1 Context and motivation

The World Health Statistics report 2017 published by the World Health Organization (WHO) listed cardiovascular diseases (CVD) as the number one cause of global deaths [2]. In 2016 alone, nearly 17.9 million people died because of CVD, and roughly 85% of these deaths were attributed to heart attack and stroke [2]. Cardiac arrhythmias comprise a large class of CVD, and are associated with an increased risk of heart attack and stroke. Among all types of arrhythmias, atrial fibrillation (AF) is most prevalent (affecting nearly 1% of the world's population); patients with AF have a six-fold higher chance of strokes, and their mortality rate is twice that of the patient in normal sinus rhythm (NSR) [67, 84, 171, 204, 186]. In addition, the financial impact of untreated arrhythmias leading to hospitalization is enormous. According to one estimation, AF alone costs nearly $26 billion in US healthcare, and it is expected to rise in the coming year as the population is growing older [91, 81, 125]. Similarly, in Denmark, roughly €73–95 million in annual healthcare cost is attributed to AF [82]. Early diagnosis of arrhythmias holds the key to prevent heart attacks and further complications [146, 41].

Electrocardiogram (ECG) analysis is one of the most economical, simple, and non-invasive ways to detect arrhythmias [5, 57, 142]. However, due to their sporadic and often asymptomatic nature, it is difficult to diagnose arrhythmias (in the early stages) during routine hospital ECG examinations. Even with a week-long Holter monitoring, nearly 30% of AF episodes remain undetected [94]; thus, more extended continuous ECG monitoring in a patient's free-living conditions is usually required [75, 155].

Traditional wired ECG Holter monitors limit the possibility of longitudinal monitoring in the patient's natural setting, as they are bulky and restrain the patient's movements [85, 159]. They are also restricted by low storage capacity and short battery life. However, recent advancements in mobile and wearable ECG technology has opened a new frontier for preemptive arrhythmia diagnosis [151, 93, 41, 120] since

such mobile and wearable ECG devices allow the collection of longitudinal ECG in the patient's natural settings [168].

Even though continuous monitoring can help in the early diagnosis of arrhythmias, it produces several challenges. First, longitudinal self-monitoring in a patient's free-living conditions suffers from lack of sustained patient engagement [147], poor signal quality [144, 147, 45], and recall bias in the patient's self-reported symptoms and activities diary [184, 142, 202]. During the ambulatory continuous Holter monitoring, patients are required to maintain a diary to note down unusual symptoms and activities. The clinicians use the diary information to map the reported symptoms with the ECG during the ECG examination. However, the process of keeping the symptoms diary suffers from frequent non-compliance, which in turn further limits the diagnostic value of the ambulatory Holter monitoring [202].

Second, inspecting and analyzing the enormous amount of data produced by continuous monitoring is tedious and resource-consuming [159]. To support the automatic detection of arrhythmias, numerous computer-aided algorithms have been proposed [40, 136, 74]. These algorithms have evolved from traditional feature engineering (FE)-based techniques which required signal pre-processing, handcrafted feature extraction, and classification to more recent end-to-end deep learning (DL)-based algorithms [40, 10, 136]. The DL-based arrhythmia detection algorithms in particular have shown a high level of performance improvement in their diagnostic capability [10, 141, 143, 48, 181, 65, 191, 111, 193, 29]. Despite these advancements, bringing these algorithms to widespread adoption for a real-time diagnosis still remains an open challenge [28, 40, 43].

Third, the majority of these algorithms have been trained and evaluated on open access datasets from PhysioNet [59]. These public datasets are primarily collected in the controlled clinical environment and are relatively clean [55]. When algorithms trained and evaluated on these datasets are applied to patient-operated ECG from ambulatory free-living conditions, they result in performance degradation, and a high false positive rate (FPR) [64, 135, 46]. This is partly due to the low signal quality of patient-operated ECG from ambulatory free-living conditions as compared to the conventional ECG recordings [172, 64]. In addition, in ambulatory ECG under free-living conditions, artifacts mimicking arrhythmias and causing misclassification are a well-known challenge [116, 51, 114, 138].

Fourth, FPR, even if small, can lead to over-diagnosis and patient anxiety [28, 96]. Even in the clinical setting's mixed screening mode where cardiologists manually verify the arrhythmia detection algorithm output, the high FPR in longitudinal screening increases cardiologists' workloads. It takes away their focus and time from treatment to diagnostics. Therefore, for the mobile and wearable technology to enable continuous screening of arrhythmias in patients' natural settings, this FPR problem needs to be addressed.

Fifth, for performance improvement so far, automated arrhythmia detection algorithms have mainly relied on ECG morphology [74]. However, under free-living conditions, contexts such as food intake, body posture changes, and physical activity play a vital role, as ECG morphology changes with these contexts [178, 40].

Computer-aided arrhythmia analysis without understanding the ambulatory context is prone to misdiagnosis and high FPR [40, 43].

In this dissertation, we explore the use of context-awareness to improve the arrhythmia analysis and reduce the FPR on longitudinal patient-operated continuous ECG under free-living conditions. The concept of context and context awareness has different notions in various areas of computer science [12]. However, in this thesis, the term context refers to the information about one or more aspects of the patient's ambulatory environment, such as physical activities, body positions, whereabouts, gender, age, food intake, lifestyle, mental state, and unusual symptoms/events. In the dissertation, the terms AF and arrhythmia will be used interchangeably, while among all types of arrhythmias, AF has been our primary focus.

1.2 Problem statement

For improving the performance of ambulatory arrhythmia detection, automated arrhythmia detection algorithms have mainly relied on ECG morphology [40]. During the manual review, it relies on the expertise of cardiologists to differentiate between the ECG morphology change due to heart conditions and the confounding artifacts due to movement/motion. However, such methods are prone to misclassification and over-diagnosis [40, 11, 23]. Under free-living ambulatory conditions, cardiovascular variability and ECG interpretation can vary according to the patient's physical contexts such as activities, stress, food intake, or clinical contexts (comorbidities like diabetes, hypertension) [169, 178]. For instance, an ECG showing a heart rate (HR)$\geq$100 BPM at rest would be considered tachyarrhythmia [13], whereas, during exercise or physical work, it would be considered normal.

Recent reviews of computer-aided arrhythmia detection techniques by Pudukotai et al. [40] and Zahra et al. [43] highlight that, under free-living conditions, without understanding the patient's context in which the ECG was collected, automatic arrhythmia analysis remains prone to misclassification [40, 43]. The importance of incorporating the patient's context in the ECG analysis algorithms is also highlighted by Smulyan [169]. Even during a manual analysis of short in-hospital ECG, whenever cardiologists find the ECG snippet inconclusive, they rely on knowledge about the patient's extensive context or about arrhythmia epidemiology to make a better assessment [65].

Prior studies have shown that even though the onset and offset of arrhythmia like AF are random, they follow some temporal pattern (clustered in the morning hours and the evening/late-night) and are associated with some external triggers like stress level, food intake, and physical efforts [66, 179]. Such information about the arrhythmia triggering and FPR-prone contexts can be combined to dynamically fine-tune the algorithm's sensitivity and specificity around those contexts.

As argued above, arrhythmia detection in ambulatory settings without taking account of the patient's ambulatory contexts is bound to result in a high FPR and over-diagnosis, especially in longitudinal screening. The use of context-awareness in

ambulatory cardiac condition monitoring has been widely discussed [133]. However, its utility in improving arrhythmia diagnosis remains unexplored. In this thesis, we investigate whether context-awareness can hold the key to reducing the FPR and improving arrhythmia detection performance under free-living conditions. We define the main research question of the dissertation as follows:

> **RQ:** *Can contextualized ECG data collected under free-living conditions help improve ambulatory arrhythmia monitoring and diagnosis?*

In our attempt to address this research question, we further identify a set of sub-questions. First, we need to understand what context information about the patient's ambulatory setting is relevant for improving arrhythmia analysis under such free-living conditions.

Secondly, as mentioned earlier, the longitudinal patient-operated ECG monitoring for arrhythmia screening suffers from the lack of sustained patient engagement, the recall bias in patient self-reported diary data, and poor quality ECG signals [147, 185, 202]. The poor quality ECG profoundly reduces the diagnostic quality during manual and computer-aided automatic analysis [148, 126]. For collecting the contextualised ECG, a few context-aware ECG collection systems have been introduced [164, 100, 109, 123]. However, the context collection in these systems is limited to activities or locations. The traditional wired Holter monitors are inadequate for longitudinal screening as they are bulky, and without any feedback to the patient, they remain a black box [85, 159]. As an alternative, mobile and wearable-based ECG are gaining traction for patient-operated ECG recordings under free-living conditions [162, 118]. Alongside the ECG, they are also adequate to collect the rich context information under ambulatory free-living conditions [133].

To address the above-mentioned challenges in longitudinal arrhythmia screening and to collect the contextualized ECG under free-living conditions, it is important to understand the technical and user experience (UX) design and the clinical feasibility of such a patient-operated mobile/wearable-based ambulatory contextualized ECG collection system. Therefore, we define the RQ 1 as:

> **RQ 1:** *What contextual information is relevant to collect during ambulatory ECG monitoring for improving arrhythmia diagnosis, and what is the design of mobile health technology for collecting such data from patients under free-living conditions?*

Furthermore, in general, it has been observed that ECG classification and arrhythmia detection models that show excellent performance on public benchmark datasets give more FPs when applied to ambulatory patient-operated single channel ECG [55, 64, 46]. Although it is known that patient's ambulatory contexts add high noise and confounding motion artifacts and cause high FPR in arrhythmia detection algorithms [46], whether any specific correlations exist between the patient's different ambulatory contexts and FP occurrences in the state-of-the-art arrhythmia detection

algorithm is yet to be explored. For instance, it might be the case that a particular type of activity, body position, time, place, or food intake induces more false positive (FP)s in an arrhythmia detection algorithm. A preliminary study in this direction was reported by Noh et al. [130]; they observed that a particular walking pattern (walking on a slope) was inducing a higher FPR in heartbeat detection algorithms than walking on flat surfaces or sitting. Similarly, if specific ambulatory contexts are found to be inducing more FPs, in state-of-the-art arrhythmia detection algorithms, such information can be used to dynamically fine-tune the algorithm's sensitivity and specificity on those FP-prone contexts. Therefore, to understand the influence of ambulatory context on the FPR of a state-of-the-art arrhythmia detection algorithm, we define the next research question as:

> **RQ 2:** *What is the impact of ambulatory contexts on FPR in a state-of-the-art arrhythmia detection algorithm when applied to ECG data collected under free-living conditions?*

If RQ 2 identifies that specific ambulatory contexts are found to influence FPR, can such contextual information be used to improve arrhythmia detection performance and reduce the FPR under free-living conditions? To explore this, we defined the final research question as:

> **RQ 3:** *How can arrhythmia detection algorithms be improved by using contextual information obtained under free-living conditions?*

1.3 Research methods

Following the recommendations from Bardram's 'Fish Model' [14], this dissertation's research methods were a combination of both empirical and theoretical works. Mackay and Fayard's [113] 'Triangulation Model' was used to categorise the variety of activities in each task. An overview of the activities performed at the level of theory, design, and observation is presented in Figure 1.1.

Methods to answer RQ 1: First, a literature review and interviews with the cardiologists were conducted to understand which patient contextual information during the ambulatory screening period could be relevant for improving arrhythmia diagnosis. Thereafter, to design a system for collecting such contextualized ECG, we followed a user-centered design (UCD) approach [63] and applied the patient-clinician-designer (PCD) framework [115]. The UCD method attempts to find a suitable design trade-off by examining diverse and seldom contrasting interests from the perspectives of all three stakeholders, namely the patient, the clinician, and the designer. The PCD framework gives a structured means for reconciling co-design activities to find proper design solutions. The co-design activities spanned 16 months (from April 2018 to July 2019) and involved six patients, and four clinicians. The proposed contextualized ECG collection system is named *mCardia*.

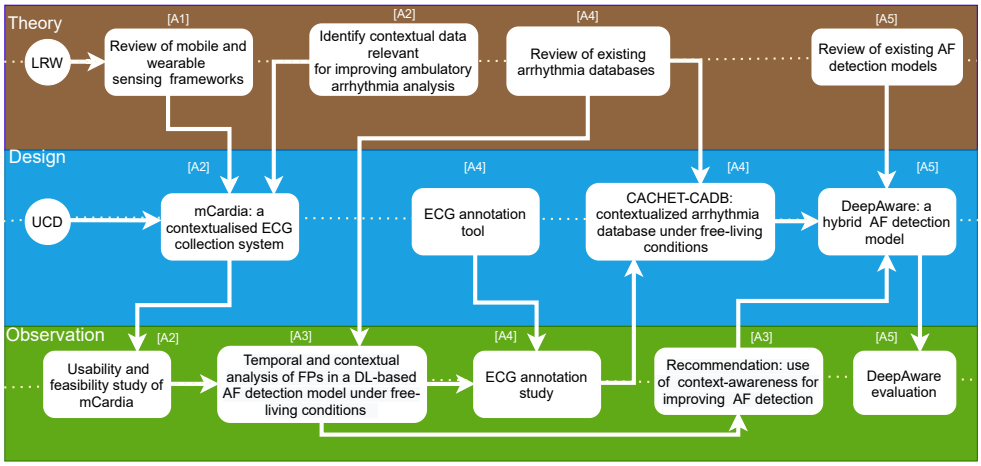


Figure 1.1: An outline of the dissertation. The color-coding separates the activities performed at each level: Theory, Design, and Observation. Each activity in the rectangle is annotated with the article number in which it is described. UCD: User-centered design, LRW: Literature review.

After completing the design process, the next task was to identify a suitable framework that could help build the *mCardia* system. In recent years many generic mobile and wearable sensing frameworks have been developed for assisting in building new mobile health (mHealth) sensing applications without reinventing the wheel. To identify a suitable framework for building the *mCardia* system, a systematic review of mobile and wearable sensing frameworks was conducted. The review revealed that most generic mobile and wearable sensing frameworks were not maintained and adequately documented after the initial releases. More importantly, they did not support a single language code-base for building a cross-platform (iOS and Android) application. Therefore, we opted to use our in-house CARP Mobile Sensing (CAMS) framework [17, 18] and the Research Package framework [73] for implementing the proposed *mCardia* system. The further implementation and architectural details of *mCardia* are described in chapter 3.

Thereafter, a study was conducted to evaluate the usability and clinical feasibility of the *mCardia* system under free-living conditions. We recruited 33 patients already diagnosed with or suspected of having arrhythmia in both India and Denmark. Patients used the *mCardia* system for a minimum of two weeks in their free-living conditions. A mixed-methods research [83] approach was employed for analyzing the feasibility study's data. The quantitative analysis consisted of analyzing usage patterns, system yield, and signal quality analysis of collected ECG data. The qualitative analysis consisted of analyzing the end of study semi-structured interviews and responses of the CACHET Unified Methodology for Assessment of Clinical Feasibility (CUMACF) [16, 19] usability questionnaires. The comprehensive HCI-based evalua-

tion gave us insights into usage patterns, the perceived usefulness of such technology for longitudinal arrhythmia screening, the system's technical feasibility in the wild, and the associated challenges [153]. Two clinical case studies were conducted to investigate the clinical usefulness of collected contextualized ECG in the arrhythmia screening process. The case studies demonstrated how and where contextual data helped the cardiologist make better assessments of patient-reported symptoms during the manual examination of the ECG. Further details of the feasibility study are described in chapter 3.

Methods to answer RQ 2: Through RQ 2, we were interested in identifying if any correlations existed between the patient's ambulatory contexts and FPR occurrences. If specific user contexts might be inducing more FPs in a state-of-the-art DL-based AF detection algorithm under free-living conditions, then such information might be used to dynamically fine-tune the algorithm's sensitivity and specificity on those FP-prone contexts. To address RQ 2, both quantitative and observational research methods are applied. First, a state-of-the-art, end-to-end, DL-based AF detection model is built and tested on public benchmark datasets. Standard metrics, namely sensitivity (Se), specificity (Sp), accuracy (Acc), and FPR, were used for evaluation and comparison of its performance on benchmark arrhythmia datasets. They are defined as follows:

$$Se = \frac{TP}{TP + FN}, \tag{1.1}$$

$$Sp = \frac{TN}{TN + FP}, \tag{1.2}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1.3}$$

$$FPR = \frac{FP}{FP + TN}, \tag{1.4}$$

where TP, FN, and TN stand for true positives, false negatives, and true negatives, respectively. After ensuring that the model has state-of-the-art performance on the public benchmark datasets, it was then applied to contextualized ECG data from free-living conditions. Further details of the DL-based AF detection model used in this investigation are described in chapter 5. A total of 215 days of data from 21 arrhythmia patients (collected during the *mCardia*'s feasibility study in RQ 1) was used. The model's output was manually examined. A mobile ECG annotation tool was designed to facilitate the manual annotations process. First, a biomedical engineer annotated the out rightly noisy or NSR segments. All the ambiguous segments were sent to two cardiologists for annotations via the mobile ECG annotation tool. After obtaining the ground truth via the manual annotation process, the true positive (TP) and FP segments marked by the arrhythmia detection model were plotted against the patient context, such as activity, body position, movement acceleration, and stress. Through the visual inspection of each 24 hour-long data section, the

correlations between the patient's ambulatory contexts and TP/FP marked by the DL-model were ascertained.

The ECG data collected during the *mCardia*'s feasibility study and the manual ECG annotations process culminated in the design and development of a 259-day-long contextualized arrhythmia database from free-living conditions. The database is named CACHET Contextualised Arrhythmia Database (CACHET-CADB). Figure 1.2 shows the procedure followed in the design and development of the dataset. The DataAnalyzer Tool [58] was used for processing the 3D accelerometer, rotation rate, and pressure sensor data from the collected ECG data. At first, AF detection models classified and labeled the rhythms, which were then manually annotated by two independent cardiologists using the ECG annotation tool. We followed a 100% inter-rater agreement policy between the two cardiologists. Samples with diverting annotations were discarded. Also, a cross-correlation-based signal quality assessment was done to ascertain the amount of noisy data in the database. Further details for CACHET-CADB's design and development are presented in chapter 6 and article [A.4].

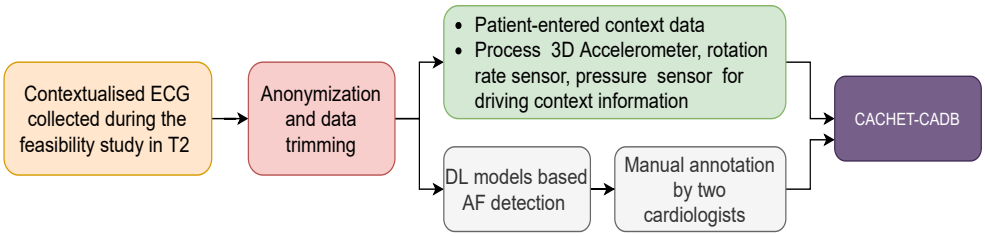


Figure 1.2: Major tasks performed to design and develop CACHET-CADB

Methods to answer RQ 3: For performance improvements a post-processing heuristics approach has been previously suggested in the literature [149, 137, 65]. In the investigation of RQ 2, we found that three specific ambulatory contexts: change in body position, activity change, and sudden movement acceleration, induced 62% of non-trivial FPs in a state-of-the-art AF classification algorithm when applied on contextualized ECG collected under free-living conditions. These findings were used to build a context-aware-heuristics module that checks if context change has been detected in a given or the previous input window to adjust the final output of the model dynamically. This context-aware-heuristics module was combined with two other DL-based models in the post-processing stage. The combined hybrid model for AF detection is named *DeepAware*. The CACHET-CADB was used for testing the context-aware-heuristics module and the overall performance of the combined model.

The standard metrics of Se, Sp, Acc, FPR, and CM were used for evaluating the performance of the proposed model. Its generality was tested on five datasets (three public and two private). To compare the performance of *DeepAware* with the state-of-the-art DL models on public datasets, we also conducted a literature review of

DL-based AF detection models. The comparison with the public arrhythmia dataset was made by disabling the context-aware-heuristics module as they do not contain context data. The same standard metrics of Se, Sp, Acc, FPR, and CM were used for performance comparison. Further details of methods are described in chapter 7 and the article [A.5].

1.4 Research contribution

The research contributions of this thesis are as follows:

1. **A systematic review of mobile and wearable sensing frameworks:** To find a suitable framework for building *mCardia*, a systematic review of available mobile and wearable sensing frameworks was performed. The review provides an overview of the state-of-the-art in mHealth sensing frameworks and proposes new features and functionality needed in future generic mHealth sensing frameworks. The review can help researchers and developers identify which framework is appropriate for building mHealth applications.

2. **Design, implementation, and evaluation of a contextualized ECG collection system for longitudinal arrhythmia screening:** We identified the challenges in longitudinal arrhythmia screening and the relevant contextual information under free-living conditions that could help improve ambulatory arrhythmia diagnosis. A contextualized ECG collection system, *mCardia*, for longitudinal arrhythmia screening under free-living conditions has been proposed. Through its device-independent plugin-based software architecture, it allows integration with any new ECG devices without modifying the application. Furthermore, a study was conducted to evaluate *mCardia*'s technical robustness, usability, and clinical feasibility via two field deployments, and over 8000 hours of contextualized ECG data has been collected. Two clinical case studies were conducted to assess the clinical usefulness of the collected contextualized ECG data. These studies demonstrated how and where the collected contextual data helped cardiologists make a better assessment during the manual analysis of ambulatory ECG.

3. **Investigation into the contextual and temporal distribution of false positives in a deep learning-based atrial fibrillation detection algorithm:** An experiment was conducted to understand the impact of ambulatory contexts on FPR in a state-of-the-art AF detection algorithm under free-living conditions. We identified the free-living ambulatory contexts that induced the non-trivial FPs in an AF detection algorithm, which otherwise has an excellent performance on benchmark arrhythmia datasets. The information about these FP-inducing contexts in free-living ambulatory settings can help dynamically fine-tune the algorithm's sensitivity and specificity around those FP-prone con-

texts. Based on these findings, we provide the design implication for future DL models to improve the FPR under free-living conditions.

4. **Design and development of a contextualized wearable ECG dataset for arrhythmia classification under free-living conditions:** A contextualized ECG database named CACHET-CADB was developed and made publicly available. It contains 259 days of contextualized ECG recordings from arrhythmia patients in free-living conditions. It provides 1602 ten-second-long manually annotated samples of ECG belonging to AF, normal sinus rhythm (NSR), Noise, and 'Other' (all other) rhythm types. Compared to existing public datasets, CACHET-CADB contains the patients' ambulatory contextual information, which, if incorporated in arrhythmia detection model designs, can improve their performance for free-living conditions. CACHET-CADB can help the medical research and ML-community to build and evaluate arrhythmia detection models that can work in patient-operated mobile and wearable ECG devices under free-living conditions.

5. **Design and development of *DeepAware*–a hybrid algorithm for AF detection using deep learning and context-aware heuristics:** We designed, developed, and evaluated *DeepAware*, which is a hybrid algorithm for AF detection using deep learning and context-aware heuristics. The proposed *DeepAware* algorithm is more generalized and beats the state-of-the-art on multiple public benchmark arrhythmia datasets. It also demonstrates how the information about the patient's ambulatory contexts under their free-living conditions can significantly reduce the non-trivial FPs.

1.5 Scientific publications in thesis

In Figure 1.1, different activities that constitute this thesis are annotated with their corresponding scientific publications. This section briefly outlines these key publications.

Mobile and Wearable Sensing Frameworks for mHealth Studies and Applications: A Systematic Review. In: *ACM Transactions on Computing for Healthcare 2 (1), 1-28*, 2020 [A.1]

A review of generic mobile and wearable sensing frameworks was conducted to find a suitable framework for building a contextualized ECG collection system. This article presents the results of that systematic literature review. It provides a comprehensive analysis of functional and non-functional features supported by existing frameworks,

the stakeholders they support, and the type of health studies in which they were used. It also offers new recommendations for future generic mHealth sensing frameworks.

mCardia: An Ambulatory Context-Aware ECG Collection System for Arrhythmia Screening In:[Accepted] *ACM Transactions on Computing for Healthcare*, 2021 [A.2]

The paper on the *mCardia* system explains its design, development, usability, and clinical feasibility via a field deployment study. The feasibility evaluation was a mixture of qualitative and quantitative results.

Contextual and Temporal Distribution of False Positives in a Deep Learning Based Atrial Fibrillation Detection Algorithm: An Investigation. In: [Submitted] *Expert Systems with Applications*, 2021 [A.3]

This article investigates the contextual and temporal distribution of false positives in a deep learning-based AF detection algorithm. The algorithm was first trained and tested on public datasets and showed state-of-the-art performance on training and other public datasets. After that, an investigation was done to find the user contexts responsible for inducing more FPR when applied to ECG from free-living conditions.

ACHET-CADB: A Contextualized Ambulatory Electrocardiography Arrhythmia Dataset In:[Submitted] *Scientific Data - Nature*, 2021 [A.4]

The paper explained the design, development, and validation process of CACHET-CADB, an ambulatory ECG database from arrhythmia patients under free-living conditions. The database contains 262 days of contextualized ECG and 1602 10 second manually annotated ECG segments of AF, NSR, Noise, and 'Other' rhythm types. This database is a rich source for validating the AF detection model's performance on patient-operated ECG from free-living conditions. Contextualized understanding of the arrhythmia detection algorithms' output can make them more transparent and help identify the error source in end-to-end classification algorithms.

DeepAware: A Hybrid Deep Learning and Context-Aware Heuristics Based Model for Atrial Fibrillation Detection. In: [Manuscript in preparation] [A.5]

This paper on *DeepAware* presents the design, development, and evaluation of a hybrid AF detection model using deep learning and context-aware heuristics. It achieves state-of-the-art performance on public datasets and also significantly reduces the FPR (while maintaining high sensitivity) on ECG data under free-living conditions.

1.6 Thesis overview

The outline of the thesis is as follows. Chapter 2 provides background information on the cardiac physiology associated with heart arrhythmias, contextualized ECG collection systems, and the public databases used for training and evaluating arrhythmia detection algorithms. After that, it summarizes the state-of-the-art, DL-based arrhythmia detection models. This background and related work chapter ends with a description of research gaps. Chapter 3 presents the technical and UX design and implementation of *mCardia*–a contextualised ECG collection system. Thereafter, chapter 4 presents *mCardia*'s usability and clinical feasibility study in two field deployments. Both of these chapters are based on article [A.2]. Chapter 5 investigates how the patient's free-living ambulatory contexts influence the FPR in a state-of-the-art AF detection algorithm. Chapter 6 presents the design, development, and validation of the proposed contextualized arrhythmia database CACHET-CADB. Chapter 7 describes the design development and evaluation of the proposed *DeepAware* model and demonstrates how combining context-aware heuristics with DL can significantly reduce the FPR under free-living conditions. In chapter 8, we discuss this research's findings in light of the main research questions (**RQ**s). This chapter also outlines some of the core limitations of the present research and suggests pointers for future work.

CHAPTER 2

Background and Related Work

This chapter first describes the necessary medical background to understand this thesis. After that, the current state-of-the-art in ambulatory ECG monitoring, benchmark arrhythmia datasets, and DL-based arrhythmia detection algorithms are described. Towards the end, we summarize the research gaps and provide an overview of how these gaps are addressed in this thesis.

2.1 Medical background

In this section we present the brief medical background of arrhythmia and its associated terminologies.

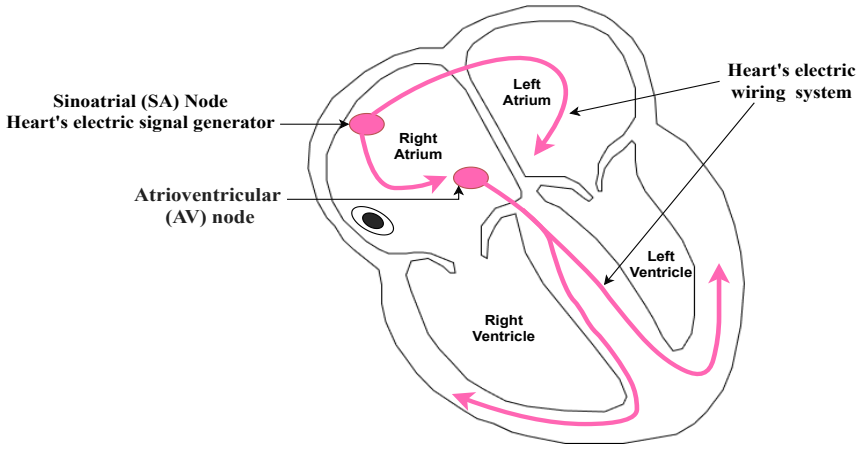


Figure 2.1: Cardiac conduction system of the heart

2.1.1 Electricity of the heart

Figure 2.1 shows an overview of the cardiac conduction system of the heart. The human heart is a muscular pump that supplies the blood flow in the body. An electrical impulse regulates this pumping mechanism. It starts in the sinus atrial (SA) node and then propagates to the left and right atria via atrial myocardial cells [86]. These electrical impulses of the heart can be recorded by putting electrodes on the chest. The recorded impulses of the heart are called an electrocardiogram (ECG). Any hindrance in the heart's electrical pathways can create an abnormality in its natural pumping mechanism/rhythms.

2.1.2 What are arrhythmias and how are they diagnosed?

Arrhythmia refers to a group of conditions associated with heart rhythm where the heartbeat becomes irregular due to a change in the normal sequence of electrical impulses of the heart [1]. Based on the type of effect they have on the heart rate/rhythm, they are categorized into several types, such as bradycardia (heart rate too slow), tachycardia (too fast) and atrial fibrillation (upper heart chambers contract irregularly) [1]. Also, according to their origin in the heart (i.e., ventricular or atrial), they can be classified as ventricular arrhythmias or supraventricular arrhythmias. Among the different types of arrhythmias, atrial fibrillation (AF) is the most prevalent and the leading cause of stroke in elderly patients [21, 167]. Early detection of such arrhythmias can help physicians effectively manage them with anticoagulant medications and so reduce the risk of further complications [71].

The electrocardiogram (ECG) analysis is the easiest, most economical, and non-invasive way of detecting arrythmias and other cardiac problems [150, 40]. However, early diagnosis of arrhythmias using ECG is challenging in many ways, since many arrhythmias remain asymptomatic and might not show up during the routine, short, in-hospital ECG. In such cases, it requires long-term monitoring ambulatory ECG for 24 hours or longer using Holter monitors.

2.1.3 The ECG and its different components

The ECG is a time series signal representing the heart's electrical activity, with a few millivolts amplitude and a frequency range of 0.01–250 Hz [182, 110]. As shown in Fig 2.2, the ECG corresponding to a single cardiac cycle contains 5-major components: P, Q, R, S, and T. They originate in different part of the heart and they are indicative of the heart's functioning. In the single heartbeat, the origin of each of these five components is as follows [199, 40, 24, 38].

- **P-wave**: The P-wave is produced during the depolarization of the left and right atria. It represents the time required for an electrical vector from the sino-atrial node (SAN) to spread throughout the atrial musculature. During this process, an electrical vector originating in the SAN spreads throughout the

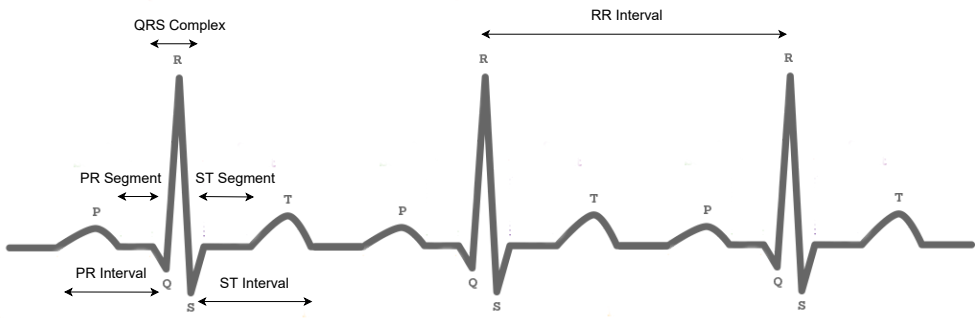


Figure 2.2: Various components of an ECG signal

atrial musculature. The normal P-wave is usually $\leq$2.5 mm tall and $\leq$120 ms in width.

- **QRS complex**: The QRS complex comprises the Q-wave, R-wave, and S-wave and indicates the right and left ventricles' rapid depolarization. The Q-wave is the first portion of the ventricular depolarization if it is negative, whereas the first upright deflection is the R-wave. The subsequent negative deflection is the S-wave. The QRS complex has a much larger amplitude than the P and T-waves.

- **T component**: The T-wave is a positive deflection after every QRS complex and represents the ventricles' repolarization phase. The ST segment is measured from the end of QRS to the start of the T-wave and represents the interval between ventricular depolarization and repolarization.

These morphological characteristics (e.g., the amplitudes, inter- and intra-heartbeat intervals of each of these five waves) of the ECG signal are the standard features used to detect arrhythmias. For instance, the RR intervals (shown in Fig 2.2) between the consecutive heartbeats can indicate the type of heart rhythm (i.e, regular or irregular). In the presence of arrhythmias like AF or sinus arrhythmia, this RR interval between heartbeats becomes irregular. Similarly, the PR segment being $\geq$0.20 s indicates the first-degree heart block [150].

2.1.4 Atrial fibrillation and ECG morphology during atrial fibrillation

As mentioned earlier, of all other types of arrhythmias, AF is the most prevalent, hazardous, and commonly under-diagnosed [56, 154]. More importantly, it is single

handedly causing an immense economic burden on health care [82] and therefore AF will remain the primary focus of this thesis.

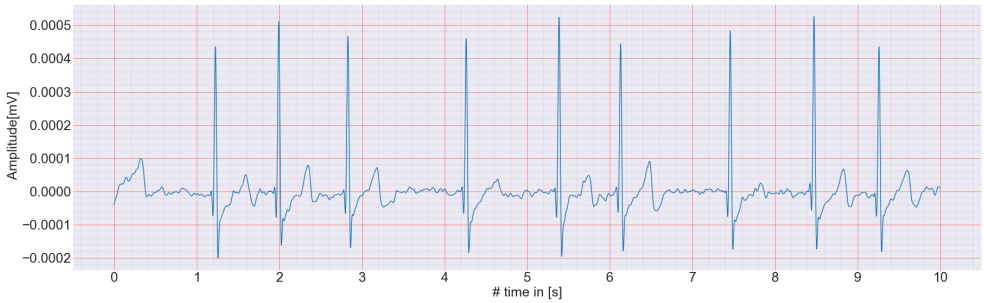


Figure 2.3: ECG morphology during the AF: (1) missing P-waves, (2) and irregular RR intervals.

AF characteristics: As shown in Figure 2.3, ECG morphology during the AF has two main characteristics:

- Heartbeat irregularity (the distance between the consecutive R-peaks)
- Absence of the P-waves

Four different types of AF conditions exist, namely (1) paroxysmal AF, (2) persistent AF, (3) long-standing persistent AF, and (4) permanent AF [32]. Among them, persistent and permanent AF conditions are easy to diagnose during an in-hospital ECG. However, the paroxysmal AF can be brief, infrequent, and asymptomatic at times and it therefore remains undetected in the routine in-hospital ECG.

2.2 ECG monitoring for arrhythmia detection

The ECG monitoring devices range from single-lead to 12-lead ECGs [160]. Based on the ECG monitoring setting, the ECG collection systems for arrhythmia analysis can be categorized into two broad categories:

- Hospital setting
- Ambulatory or free-living conditions

The in-hospital ECG monitoring systems are further divided into two: ICU types and those for non-ICU settings [26, 197, 6, 42, 176, 160]. The hospital-based arrhythmia monitoring systems are very standardized. Also, as mentioned earlier, due to their paroxysmal nature, arrhythmias are difficult to detect during routine in-hospital ECG in their early stages. Therefore, we will further focus only on the continuous ambulatory ECG monitoring systems.

2.2.1 ECG monitoring under ambulatory free-living conditions

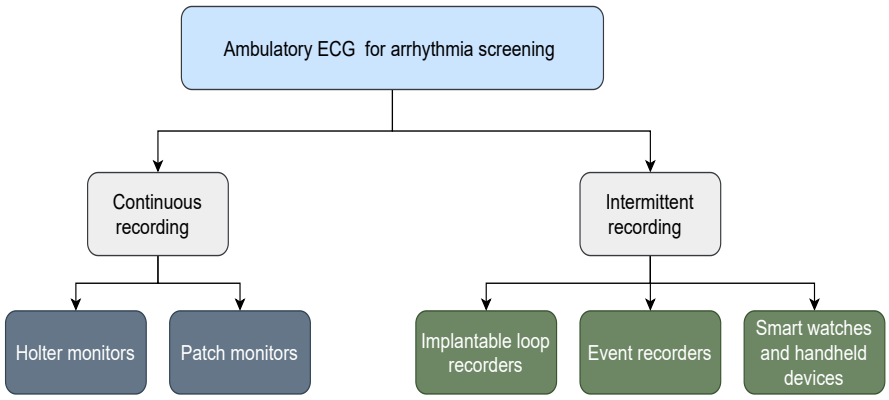


Figure 2.4: Ambulatory ECG systems based on monitoring scheme

Based on the recording scheme, ambulatory ECG monitoring systems can be categorized as continuous recorder or intermittent recorder(Figure 2.4). For continuous recording, Holter monitors and adhesive patch monitors are common. Holter monitors can usually record 2–4 channels of ECG for 24–48 hours and are therefore most suitable for patients experiencing some unusual symptoms/episodes on a daily basis [142, 54]. Traditional Holter monitors are bulky in size and, without any feedback, remain a black box for the patient [85, 159]. On the other hand, the adhesive patch monitors are usually single channel and can continuously record multiple weeks of ECG [142].

The intermittent monitoring systems include event recorders, smartwatches/handheld devices, and implantable loop recorders [142, 54, 202]. The event recorders record ECG data for a short programmable fixed amount of time (typically 1–4 minutes) when a patient experiences some unusual symptom and presses a recording button [202]. Similarly, the smartwatches and handheld devices such as AliveCor [78], Apple watch [79], and Zenicor-ECG [173] are also used for recording short (usually 30 second–1 min) ECGs. The implantable loop recorders are invasive and can record a single channel ECG for up to a year [202]. They are mostly used when the frequency of unusual symptoms is very low (e.g., once in two months) [54, 142, 124]. Compared to other monitoring devices, their cost is also significantly higher [202].

Out of the two monitoring schemes, continuous long-term monitoring is more desirable as intermittent ECG might miss asymptomatic arrhythmias [54].

2.2.2 Other modality of ambulatory arrhythmia monitoring

Apart from standard ECG-based systems, in recent years photoplethysmography (PPG) has also gained traction for ambulatory arrhythmia detection [28, 163, 101, 9,

161]. The PPG base arrhythmia detection systems collect PPG data from two sources: (1) wrist-worn wearable smart-watches or fitness trackers, (2) smartphone cameras. Due to their form factor and easy-to-use nature, the wrist-worn PPG wearables such as health trackers and watches are much more suitable for long-term arrhythmia monitoring [20, 44]. However, as reported by Dörr et.al [41], the PPG data from wrist-worn wearables is very noisy, and significant parts of the recordings are not usable for arrhythmia analysis due to low signal quality. The quantity of unusable data increases even more when used under free-living conditions due to more confounding movement artifacts. In smartphone camera-based PPG, the pulsatile time-series recordings are obtained by placing a fingertip on the phone's camera. McManus et al. [120] built a pulse waveform analysis algorithm to detect arrhythmias and claimed to have achieved an accuracy of 95.1% for AF detection. Similarly, Cardiio Rhythm [27] also reported a sensitivity of 92.9%.

It is important to note that PPG-based arrhythmia detections system are still in the exploratory phase, and ECG still remains the standard for arrhythmia diagnosis in clinical settings. Therefore in this thesis, we are especially interested in continuous ambulatory ECG.

2.2.3 Challenges in longitudinal arrhythmia screening under free-living conditions

Figure 2.5 shows the workflow of the traditional outpatient arrhythmia screening process. During this screening period, patients are advised to keep a diary to note down any unusual symptoms they might experience [142, 202]. A cardiologist uses these notes of unusual symptoms during the manual analysis to check if the occurrence of those symptoms correlates with the patient's ECG. After the recording period, the ECG data is manually extracted from the Holter device, and the symptoms diary is handed over for examination. This approach of arrhythmia screening faces the following challenges:

- **Arrhythmia mimicking artifacts:** The ECG recording under free-living conditions gets contaminated by various motion artifacts and anomalies. These artifacts often mimic arrhythmias and other cardiac events [114, 116, 119, 138], resulting in false diagnosis. In the manual analysis, without having any information about the patient's ambulatory context, clinicians rely solely on their experience to decide whether the ECG morphology changes were due to arrhythmias or confounding motion artifacts. Therefore, the ambulatory ECG analysis (manual or computer-aided) independent of the patient's physical condition and context remains prone to misinterpretation and misclassification of arrhythmias [43, 40], especially in the younger population with low arrhythmia prevalence.
- **Lack of user engagement in longitudinal ECG collection:** The longitudinal self-monitoring in a patient's natural setting suffers from lack of sustained

patient engagement [147]. The traditional wired Holter monitors used for ambulatory ECG collection are not very efficient due to their bulky size [85] and do not provide any feedback to the patient. Even the ECG patch devices used for long-term monitoring (up to two weeks) are not very engaging and remain a black box for the patient. The lack of user engagement often leads to frequent noncompliance in the patient-reported symptoms diary, which patients are required to maintain during the monitoring period [202].

- **Poor signal quality:** In ambulatory ECG monitoring, electrodes often become non-adhesive and result in poor ECG signals. Although some Holter monitors have alarms to notify the patient of the poor signal quality, patients cannot always identify and take corrective measures to fix or change the electrodes. In addition, the ECG signal quality is also affected due to various user activities under free-living conditions [144]. Without active user engagement/feedback, the quality of the collected ECG data suffers, and often a large part of the ambulatory data remains unusable for analysis [147, 45].
- **Recall bias on patient's self-reported symptoms and events:** In clinical settings, the quality and reliability of patient-generated data without an understanding of the patient's context is a matter of concern [185]. As shown in Figure 2.5, during the ambulatory morning, patients report the unusual systems (e.g., dizziness, palpitation, shortness of breath) via a paper-based diary [142]. A cardiologist or Holter nurse uses this information during the ECG analysis to map the symptoms reported in the diary with the ECG. Completion of the paper-based diary is often based on the patient's recall as the patients do not always carry it with them. During analysis, the lack of synchronization in the timestamp and frequent noncompliance [202] makes it challenging for the cardiologist or nurse to map the symptoms to the corresponding ECG. In long-term ECG under free-living conditions, this challenge of recall bias on patient-reported symptoms is multiplied manyfold.

2.3 Systems for collecting contextualised mobile ECG

With the advancement of mobile and wearable technology, the self-monitoring mobile ECG for diagnosing cardiac arrhythmia has gained traction in the last decade [117]. To deal with some of the aforementioned challenges with longitudinal ECG collection under free-living conditions and to improve the diagnostic value of self-monitored continuous mobile ECG, many context-aware ECG collection systems have been proposed in literature [123, 109, 164, 170]. Along with the ambulatory ECG, they also collect various types of user context information, either actively by engaging the user or passively through the sensors.

In the earliest work in this direction, Shirazi et al. [164] introduced *CardioViz* for long-term context-aware ECG monitoring. *CardioViz* used a phone camera and

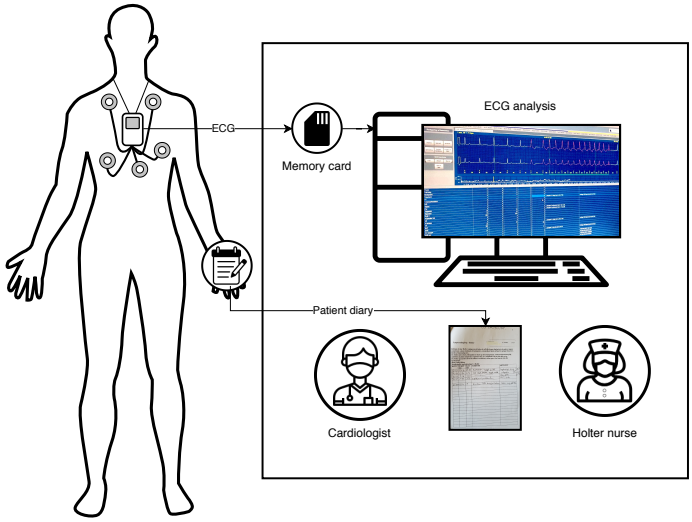


Figure 2.5: The workflow of the outpatient arrhythmia screening process (image from [A.2]).

an external GPS sensor for collecting context information and showing it together with the ECG. The system's objective was to capture the context that can hint and remind users about the situation related to a particular period in the ECG. Likewise, Belgacem et al. [22] built a mobile ECG collection system that also collected the patient location. In contrast to the above two systems in which contextual information collected during the ECG monitoring was limited to GPS location and photography, the iMote2-based system by Spadini et al. [170] collected accelerometer and environmental data such as temperature, humidity, and light intensity.

The mobile ECG monitoring system by Li et al. [109] focused on active context collection by engaging the user. Whenever any change in the ECG or heart rate was detected, it prompted an interface that asked the patient to enter the activity they were performing at that point in time. Similarly, a wearable context-aware ECG monitoring system by Miao et al. [123] also collected the user's activity; however, unlike that of Li et al. [109] it employed built-in smartphone sensors for activity recognition. They also demonstrated the usefulness of physical activity recognition to improve the arrhythmia diagnosis accuracy and identify the frequent irregular ECG patterns under different activities.

The context-aware cardiac monitoring systems by Forkan et al. [52] aimed to reduce false alerts by utilizing the context-awareness of the collected ECG data. In [53] they also introduced a context-aware system for monitoring elderly cardiac patients under free-living conditions and integrated it with social networking services for distant help and tracking by family and doctors.

Although the aforementioned systems have advanced the research in context-aware

ECG collection, the use of contextualized ECG for improving arrhythmia detection is still in infancy and needs significant improvements [133]. The existing context-aware mobile ECG collection systems have the following limitations. First, in these systems, the collected contextual data is limited; and they do not address the issue of lack of user engagement that is prevalent in longitudinal self self-monitoring [147]. Furthermore, their usability and clinical feasibility in longitudinal contextualized ECG collection under free-living conditions have not been explored. Second, they are primarily focused on passive context collection and do not collect the user-reported subjective contextual data (e.g., symptoms diary [142, 202]), which is essential during ambulatory arrhythmia monitoring.

2.4 ECG database for evaluation of arrhythmia detection algorithms

Table 2.1 provides a statistical overview of various publicly available benchmark datasets used in the evaluation of arrhythmia detection algorithms. These datasets are available on PhysioNet [59]. A detailed description of these databases is as follows:

MIT-BIH Arrhythmia Database (MITDB): The MITDB comprises 24.7-hour-long two-channel ambulatory ECG recordings collected from 47 unique participants. The ECG was recorded with an 11-bit resolution at a sampling frequency of 360 Hz. The database contains approximately 110,000 beat-by-beat annotations provided by multiple independent cardiologists. It contains 15 different types/classes of arrhythmias, of which approximately 2.16 hrs (8.16%) is AF. Every recording includes an annotation file that contains information about the arrhythmia type, its onset, R-peak location, and the type of each beat [59, 127].

MIT-BIH AF Database (AFDB): The AFDB is the most used database for developing AF detection models [40, 47, 136]. It contains 25 long-term two-channel ECG recordings of 10 hours each from patients with mostly paroxysmal atrial fibrillation. Recording took place in an in-hospital setting at Boston's Beth Israel Hospital. The ECG is digitized at a sampling rate of 250 Hz and a 12-bit resolution. The database contains four main rhythm annotation classes: (1) atrial fibrillation, (2) atrial flutter, (3) AV junctional rhythm, and (4) others. In the annotation files, these four classes are marked as AF, AFL, J, and N, respectively. Like MITDB, it also provides the beat-by-beat annotation, QRS complex, and R-peak locations. An automated detector was used to perform the beat annotation.

Open-Access Arrhythmia Database (OA-ADB): The OA-ADB [162] consists of 2000 30-second samples of sinus, atrial and ventricular arrhythmias from over 200 patients. The ECG was collected using a 6-channel wireless Holter monitor at a sampling rate of 420 Hz and 12-bit resolution. The length of the ECG recordings varied from 24 hours to 72 hours and, unlike others, it contains data from patients of a wide age range (18–85 years). The initial annotations were done by an automated

algorithm and subsequently verified by two cardiologists who manually verified the labels.

DeepQ Arrhythmia Database (DeepQ): The DeepQ [187] is another large-scale arrhythmia database that contains 897 annotated single-lead ECG recordings from 299 unique patients. Unlike MITDB and AFDB, DeepQ database only contains 15 minutes of ECG recording per patient. Each recording was done while the participant was performing three types of activity (5 minutes each): lying down, sitting, and walking. The ECG was collected using a single-lead ECG patch worn in the lead II configuration with a sampling rate of 250 Hz and 12-bit resolution. Along with the beat-by-beat annotation, it also contains rhythm episodes and heartbeat fiducial points. The annotations were done by cardiographic technicians and subsequently verified by a cardiologist. The inter-rater reliability scores are, however, not reported. Compared with existing public arrhythmia datasets from PhysioNet [59], DeepQ is more diverse and larger. DeepQ tries to mimic ECG under free-living conditions by including some common contexts such as lying down, sitting, and walking; they were collected in a controlled environment where the subjects performed these activities for 5 minutes each. In a free-living condition, user-context and confounding artifacts are not limited to just these three activities and, therefore, it does not represent patients' ECG morphologies under a truly free-living scenario.

Computing in Cardiology Challenge 2017 Dataset (CinCDB): The CinCDB [31] has 8,528 single-channel ECG samples of different lengths from 9 seconds to over 60 seconds. The data is collected using *AliveCor's* handheld ECG device (equivalent to lead I) and stored at 300 Hz with a 16-bit resolution. Unlike the other databases discussed above, recordings in CinCDB are not from continuous ambulatory monitoring. The handheld device usually records short ECG recordings up to an average of 30 seconds. During the recording, patients were advised to sit in a comfortable position without making any hand movements to avoid motion artifacts. Dataset annotation contains four types of classifications, namely NSR (59.5%), AF (8.9 %), other (28.3%), and noisy (3.3%).

MIT-BIH NSR Database (NSRDB): The MIT-BIH NSR Database (NSRDB) is an ambulatory database from 18 subjects in NSR without any arrhythmias. It is primarily used for testing the generality and FPR in the arrhythmia detection algorithms. It can provide a good estimation of a model's robustness, especially in the population with a low arrhythmia prevalence.

Others: Other databases such as CU Ventricular Tachyarrhythmia Database (CUDB) [131], MIT-BIH Noise Stress Test Database (NSTDB) [129] and the MIT-BIH Malignant Ventricular Arrhythmia Database (MVFDB) [61] also feature in arrhythmia detection literature; however, their usage was limited. The MVFDB contains 22 half-hour recordings, whereas CUDB contains 35 eight-minute ECG recordings. The annotations in both of these databases include ventricular flutter, ventricular tachycardia, and ventricular fibrillation types. NSTDB contains 15 half-hour ECG recordings, of which 3 have noise similar to that in ambulatory ECG recordings.

Table 2.1: Specifications and ECG annotations statistics of publicly available arrhythmia databases. Ch: Number of ECG channels, Freq (Hz): sampling frequency.

Database	**Ch**	**Freq (Hz)**	**Number samples**	**Sample length**	**Rhythm classes**	**Number subjects**	**Context**	**Remark**
AFDB [127]	2	250	23	10 h	4	25	✗	continuous, controlled
MITDB [128]	2	360	48	30 min	15	47	✗	continuous, controlled
NSRDB [59]	2	128	18	24 h	1	18	✗	continuous, ambulatory, controlled
DeepQ [187]	1	250	897	5 min	8	299	✗	intermittent, controlled environment
OA-ADB [162]	6	400	2000	30 s	15	200	✗	continuous, ambulatory, patient-operated
CinC2017 [31]	1	300	8528	9 s to 60 s	4	-	✗	intermittent, patient-operated
MVFDB [61, 59]	2	250	22	30 min	3	-	✗	-
NSTDB [129]	2	250	15	30 min	1	-	✗	-
CUDB [131, 59]	1	35	250	8 min	3	-	✗	-

2.4.1 Limitation of existing public arrhythmia dataset:

Although the public benchmark arrhythmia datasets described above have helped immensely to advance research into automatic arrhythmia detection, they face the following challenges:

- They are mostly collected under clinical supervision; thus, they are relatively clean and have high signal quality compared with patient-operated ECG from free-living conditions [55]. Patient-operated ECG from wearable devices is gaining traction for arrhythmia screening [162]. To enable arrhythmia detection under free-living conditions, models need to be built and evaluated on such patient-operated ECG from wearable devices rather than on clinical-grade hand-picked clean ECG (as in the case of public datasets), on which models tends to show a high performance.

- They do not provide the patient's context information during the ECG collection period; therefore, they are not suitable for context-aware analysis of classification models.

- Even though some of the databases are ambulatory, they are limited to a few hours or days. Therefore, they do not contain all the ECG morphological changes and the arrhythmia-mimicking confounding artifacts that are expected in longitudinal ambulatory free-living conditions. When classification models trained on them are applied on ECG from free-living ambulatory conditions, they show performance degradation [55].

2.5 Computer aided arrhythmia diagnosis

Although wearable Holters and ECG collection methods discussed above can help in collecting ambulatory ECG data under free-living conditions, inspecting and analyzing multi-day/week data is a resource- and time-consuming task. Over the years, numerous computer-aided algorithms have been developed to assist in the auto-detection of onset and duration of arrhythmic episodes [104, 46, 77, 191, 111, 103, 65, 33, 181, 39, 48, 193, 200]. These algorithms could facilitate and expedite the AF screening process and help in achieving early-stage diagnosis. These computer-aided arrhythmia detection algorithms can be classified into two broad categories:

1. Algorithms based on feature engineering (FE).
2. Algorithms based on end-to-end deep learning.

Each of these techniques are discussed in detail in the following sections.

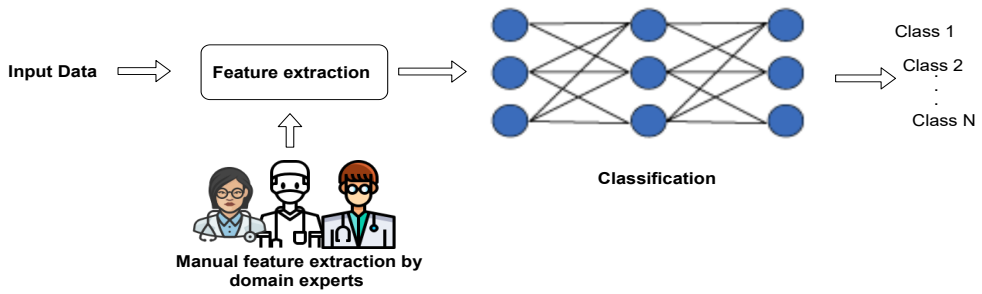


Figure 2.6: Traditional feature engineering-based approach of arrhythmia detection.

2.5.1 Algorithms based on feature engineering

Figure 2.6 shows the traditional FE-based approach of arrhythmia detection. These algorithms involve following three steps: (1) preprocessing, (2) manual feature extraction, and (3) classification. The feature extraction process uses either domain experts or conventional feature extraction algorithms. The extracted features are then fed into the classifier, such as support vector machines (SVMs) and hidden Markov models (HMMs), to produce the final ECG classification [143, 156, 203, 39, 89]. The extracted feature quality directly influences the classification performance and robustness of such algorithms. The hand-crafted feature extractions based on domain knowledge are (1) labor-intensive, (2) influenced by the expert's bias, and thus prone to errors, and (3) less robust to adaptation in the presence of any variations and noise in heterogeneous data [165, 143]. The use of conventional feature extraction algorithms such as the discrete wavelet transform, cosine transform, and discrete Fourier transform has also been explored in several articles for extracting time and frequency domain features from ECG [88, 90, 143, 158]. Numerous FE-based algorithms just on the RR intervals features have been developed in the past [37, 112, 87, 77]. Kennedy et al. [87] used RR interval features on random forests and k-nearest neighbor for AF identification. Similarly, Dash et al. [37] and Linker et al. [112] attempted to combine the atrial activity and RR intervals with AF detection.

A brief summary of traditional approaches to arrhythmia detection follows [40]:

- Among arrhythmia detectors based on frequency- or time-domain analysis, frequency-domain analysis is more powerful; however, it is computationally more expensive than time-domain.

- Traditional FE-based approaches lack stability and are susceptible to noise and confounding artifacts.

- They are widely used to detect limited (single or few) arrhythmia types, implying that they are not suitable for making a generic arrhythmia detector.

- The manual feature extraction needed in this approach is laborious and dependent on domain experts.

- The classification performances achieved by some of the automatic feature extraction approaches like filtering and auto-correlation are poor and can lead to a faulty diagnosis and health consequences.

2.5.2 Algorithms based on end-to-end deep learning

The field of deep learning (DL) has achieved remarkable success in areas like image recognition, and natural language processing [68, 34, 166, 107]. Its ability to extract complex features and recognize patterns directly from time-series data without expert intervention has gained many researchers' interest in applying them in the domain of ECG analysis and arrhythmia detection. In the last five years, deep learning for arrhythmia detection has gained significant momentum [43, 136, 40, 118]. Six deep learning architectures, namely Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multilayer Perceptron (MLP), Long-Short Term Memory (LSTM), Gated Recurrent Units (GRU), Deep Belief Network (DBN), or their combinations, are frequently used for arrhythmia detection [43, 136]. A brief overview of these deep learning architectures is provided in section 2.5.3.

Advantages of deep learning-based arrhythmia detection algorithms over the FE-based algorithms:

1. They remove the need for manual feature extraction (see Fig 2.7) and can achieve **end-to-end classification** with more robust and abstract feature extraction [10, 143].

2. Unlike FE-based algorithms, they are more robust in the presence of noise/artifacts and adapt very well to heterogeneous data [143, 40].

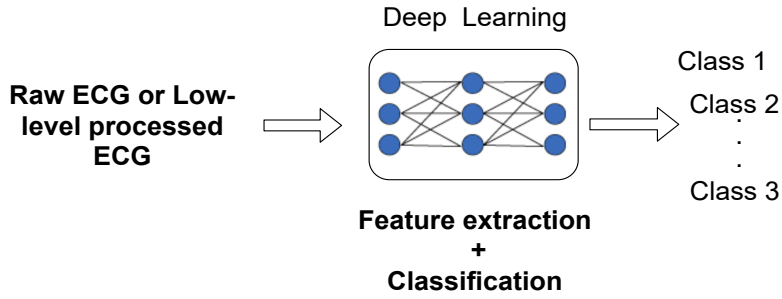


Figure 2.7: Deep learning-based end-to-end arrhythmia classification.

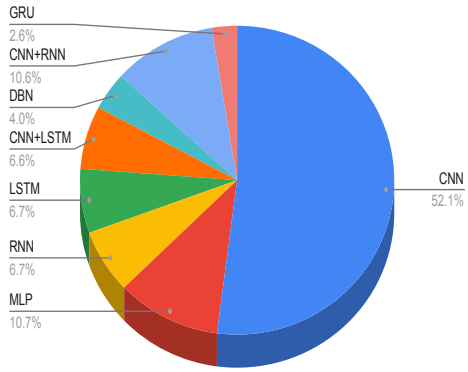


Figure 2.8: (a) Most frequent deep learning architectures used for arrhythmia classification in literature (image data source [43]).

2.5.3 Overview of deep learning architectures used in ECG classifications

This section provides an overview of the six popular DL-architectures (in Figure 2.8) used in ECG classifications.

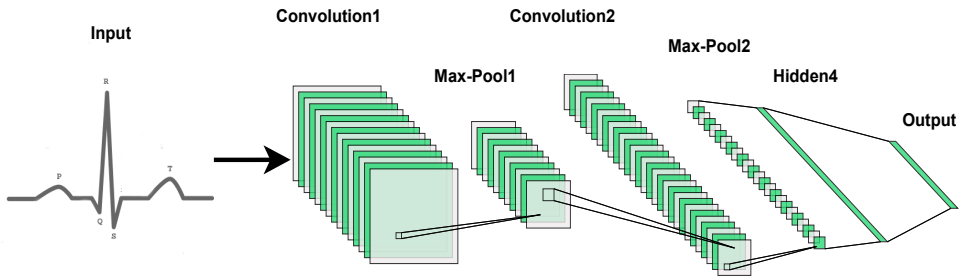


Figure 2.9: Architecture of the CNN.

Convolution neural network: The CNN [134] is a type of artificial neural network with some specialization for picking out or detecting patterns and making sense of them. A typical CNN model contains single or multiple convolution layers, rectified linear unit (ReLU), non-linearity layer, and pooling layers followed by a fully connected layer or softmax for classification. The convolution layers are the heart of the CNN. Unlike convolutional layers and fully connected layers, the non-linearity

and pooling layers do not have parameters [8].

Recurrent neural network: RNN is a type of neural network with a feedback loop that enables information to be stored and reused within the network for sequential decisions. It uses the output from the preceding element of an input sequence to inform future elements [121, 157]. This ability of RNNs makes them attractive for modeling sequential time series data like ECG. Figure 2.10 depicts the architecture of the RNN.

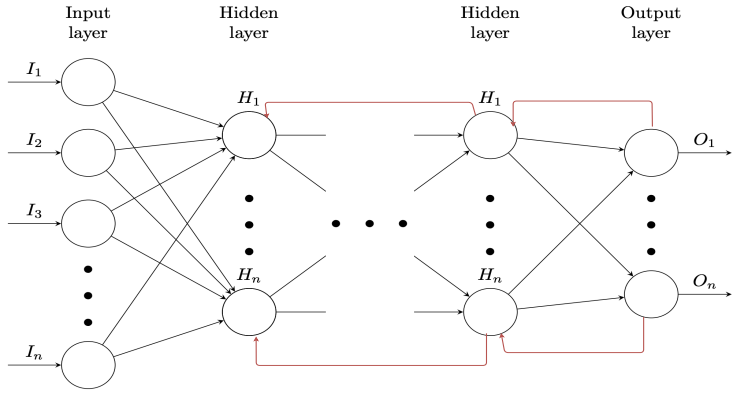


Figure 2.10: Architecture of the RNN.

Long-short-term memory: The LSTM networks [72, 60] are RNNs that can preserve information in the memory cell for a more extended period. The standard RNN architectures suffer from vanishing and exploding gradient problems and cannot hold memory for long. Using a gate control mechanism (for input and forget gates of the memory cell), LSTM can retain sequential memory for a much longer period. Among LSTM networks, bidirectional LSTM (Figure 2.11) is more popular as its memory cells can retain information from both past and future.

Deep belief network: The DBN is a DL architecture used to solve the problem of low velocity, the over-fitting phenomenon in the deep layers, and training dataset requirement [195, 7]. Figure 2.12 shows the architecture of DBN. It comprises multiple layers of 'Boltzmann Machines' and each one of them is restricted to a single visible and hidden layer [7].

Multilayer perceptron: MLP is an artificial neural network (ANN) composed of one or more layers of neurons with only forward connections to units in subsequent layers [76]. In MLP, the input layer passes the inputs to hidden layers. The nodes in the input and output layers have linear activation functions, whereas nodes in hidden layers have nonlinear activation functions with thresholds. Figure 2.13 shows the network structure of the MLP.

Gated recurrent units: GRUs are modifications to RNN's hidden layers that help it capture long-range connections and overcome the vanishing gradient problems

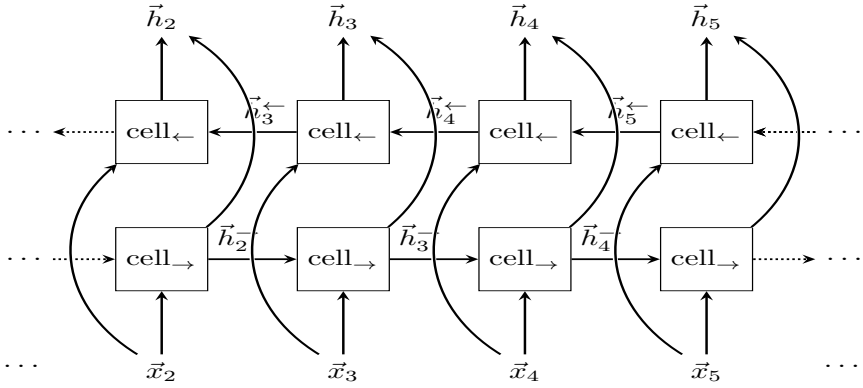


Figure 2.11: Structure of the bidirectional LSTM.

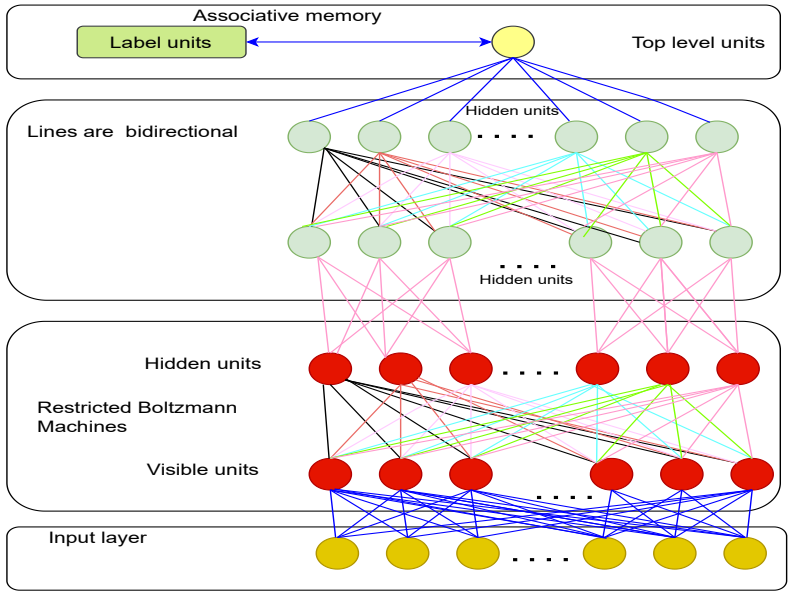


Figure 2.12: Deep Belief Network.

by using an update and a reset gate [30]. Figure 2.14 shows the architecture of Gated Recurrent Units (GRU).

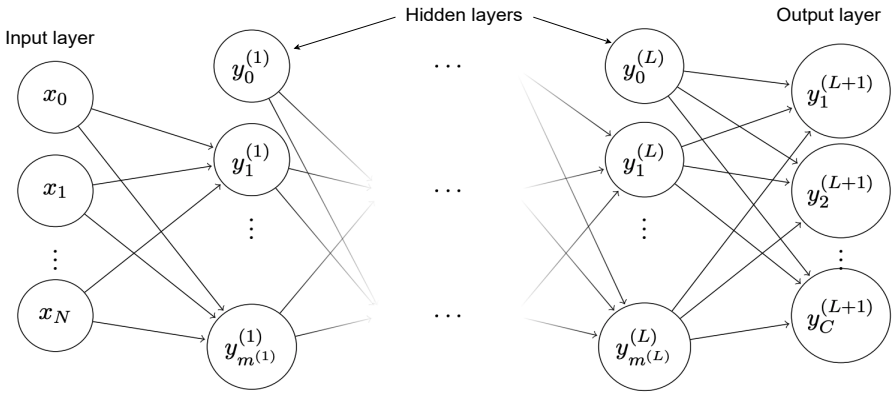


Figure 2.13: Network structure of Multilayer Perceptron with N inputs, C outputs and L Hidden layers.

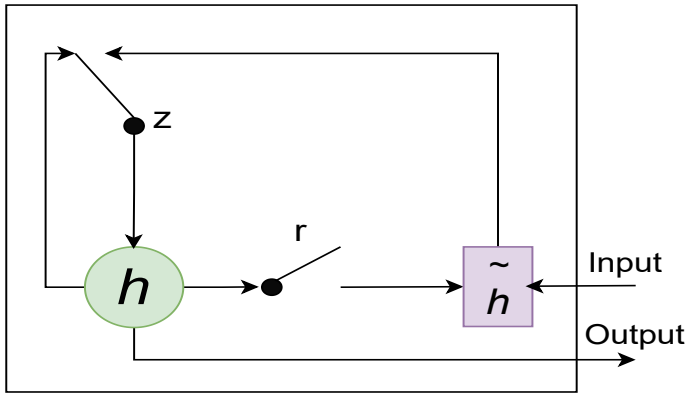


Figure 2.14: Gated Recurrent Unit. r: reset gate, z: update gates, h: activation $\widetilde{h}$: candidate activation.

2.5.3.1 State-of-the-art deep learning-based algorithms for arrhythmia detection

Table 2.2 shows the performance of various DL algorithms. The majority of these models have been trained and evaluated on datasets from PhysioNet [59]. Based on the DL approach used, feature selections, and the main emphasis of the algorithms, we have categorized the existing arrhythmia detection algorithms into the following categories.

Multi-Model approach: In a multi-model approach, one or more DL architectures were combined to achieve better classification results [10, 35, 183]. Andersen et al. [10] combined CNN and RNN to achieve an end-to-end AF detection algorithm. On five-fold cross-validation on AFDB, it achieved a sensitivity of 98.98% and a specificity of 96.95%. The generality of the model was tested on MITDB and NSRDB. The model resulted in more FPR (13.96% and 4.99%) on both of these datasets, indicating that models trained on RRI features suffer in the presence of confounding arrhythmias that resembled AF in terms of RR irregularities and ambulatory noise (in case of NSRDB). Similarly, Xiaoling et al. [183] coupled a recurrence complex network (RCN) and CNN with a majority voting methodology and achieved an accuracy of 94.59% on AFDB. In some algorithms, each sub-model was trained on different input features (e.g., RRI, raw ECG). For instance, Dang et al. [35] used CNN and a Bidirectional Long-Short Term Memory (BLSTM) based model for AF detection using RRI and heartbeat sequences (P-QRS-T waves). It comprises two BLSTM layers and two fully connected layers and achieved an accuracy of 96.59%, a sensitivity of 99.93%, and a specificity of 97.03% on AFDB.

In another multi-model approach, Salem et al. [152] proposed an algorithm based on transfer learning for arrhythmia detection by combining a densely connected CNN (DenseNet) with support vector machine (SVM). In ten-fold cross-validation on AFDB, NSRDB, MVFDB, and ESTDB, they achieved an overall accuracy of 97.23%. Furthermore, Yao et al. [193] proposed multi-scale convolutional neural networks (MCNN) trained on an instant heart rate sequence as input and achieved an overall accuracy of 98.18% on AFBD.

Multi-class classification: A multi-class approach endeavors to classify multiple classes of arrhythmias using a single classifier [65, 201, 80, 25]. Multi-class arrhythmia detection is more challenging than binary classification as the feature characteristics of many arrhythmias have a close resemblance. Hannun et al. [65] developed a DNN with 33 convolutional layers to classify 12 different types of arrhythmias from a single lead non-ambulatory ECG. The model was trained and evaluated on a single channel non-ambulatory private ECG dataset of 53,549 unique subjects. It achieved an average F1 score of 0.837. Similarly, Cao et al. [25] applied multi-scale decomposition on the residual convolutional neural network (MSResNet) on the short segment (9 seconds) ECG for classifying four types of rhythms. The derived wavelet frame (DWF) decomposition was used as input to three fast down-sampling residual convolutional neural networks (FDResNets). These FDResNets were then combined to form the final MSResNet via transfer learning. On CinCDB, the model achieved an overall accuracy of 87.12% and an average F1 score of 85.29%.

Zihlmann et al. [201] achieved an F1 scope of 82.1% on CinCDB by employing a combination of CNN and LSTM on the logarithmic spectrogram features. The one-sided spectrogram was computed from the time-domain ECG signal, and subsequently, a logarithmic transform was applied to achieve the final logarithmic spectrogram. In contrast, spectrogram features in [201], [80] was trained on RRI features and achieved an average accuracy of 88.28% for classifying NSR, AF, and AFL on a private dataset in a 10-fold validation.

R-peak/P-wave independent short segment classification: The algorithms using the RRI features usually require an input of 30–50 seconds or longer for correct assessments. However, such systems may miss out on short episodes of arrhythmia. To overcome this, several algorithms have been proposed for classifying arrhythmia from even short segments of ECG [194, 3, 189]. These algorithms usually do not rely on RRI features or a need for P-wave detection. Yıldırım et al. [194] proposed a 1D CNN-based model that takes 10 seconds of raw ECG fragment as input. It claimed to have achieved an overall accuracy of 91.33% in classifying 17 different rhythm classes from MITDB. Similarly, Wu et al. [188] used CNN with continuous wavelet transform (CWT) features of 10 seconds of ECG and achieved an overall accuracy of 97.56% on five datasets from PhysioNet [59]. Of the three wavelet transforms used, Mexh wavelet, Morl wavelet, and Cmor wavelet, Morl wavelet achieved the highest performance.

Xia et al. [189] built an AF detection algorithm by using 2D CNN on short-term Fourier transform (STFT), and stationary wavelet transform (SWT) features. This could detect AF even in a small 5-second ECG segment, whereas other RRI-based models usually require a long input window. The model was trained and tested on AFDB, and like others, its generality was also not reported. Acharya et al. [3] also built an 11-layer CNN-based arrhythmia detection model for detecting AF, atrial flutter, and ventricular fibrillation (VF). Similar to the methods of Xia et al. [189] they also permed analysis on 2-second and 5-second segments of ECG in which the accuracy on the 5-second inputs (94.9%) was slightly higher than on the 2-second (92.50%) inputs.

Person-specific classification: To make generic arrhythmia detection algorithms that can work despite all inter or intra-personal differences in the ECG, many patient-specific arrhythmia classification algorithms have been explored [69, 198]. Zhao et.al [198] used an adaptive ResNet [69] model on MITDB and achieved an overall accuracy of 98.6% for classifying five types of heartbeats. Similarly, Kiranyaz et al. [92] presented adaptive 1D CNN where CNN was first trained on common data and then on 5 minutes of patient-specific data on MITDB.

Table 2.2: R-class: No of rhythm/arrhythmias classified, STFT: Short-term Fourier transform, SWT: Stationary wavelet transform, AFL: Atrial flutter, SR: sinus rhythm, DWF: Derived wavelet frames, LS: Logarithmic spectrogram, IHR: Instant heart rate sequence.

Models	**Features**	**Performance**				**Database**	**R-class**	**Ref**
		F1-score	Se	Sp	Acc			
LSTM, CNN	RRI	-	98.98%	96.95%	97.80%	AFDB, MITDB, NSRDB	AF	[10]
LSTM	RRI	-	98.32%	98.67%	98.51%	AFDB	AF	[49]
CNN	STFT, SWT	-	98.79%	97.87%	98.63%	AFDB	AF	[189]
CNN	Raw ECG	-	98.09-99.13%	81.44-93.13%	92.5-94.9%	AFDB, MITDB, CUDB	AF, VF AFL	[3]
1D CNN	Raw ECG	-	64.4-95.9%	98.1-99.5%	96.6-99%	MITDB	5	[92]
CNN	Raw ECG	-	95.32%	91.04%	93.18%	MITDB, MVFDB, CUDB	VT,VF	[4]
DNN	Raw ECG	83.7%	-	-	-	Private	12	[65]
2D CNN	CWT	-	99.41%	98.91%	99.23%	AFDB	AF	[70]
CNN, BLSTM	RRI, Heartbeat sequences	-	99.93%	97.03%	96.59%	AFDB	AF	[35]
CNN	MFSWT	-	74.96%	86.41%	81.07%	AFDB	AF	[192]
LSTM, CNN	RRI	-	-	-	88.28 %	Private	AF, AFL, NSR	[80]
CNN	RRI	-	-	-	93.6%	MITDB	7	[145]
MSResNet	DWF	85.29%	-	-	87.12%	CinCDB	AF, NSR, Noise, Other	[25]
CNN	CWT	-	97.56%	99.19%	97.56%	MITDB, AFDB, MVFDB, NSRDB NSTDB, LTAFDB	AF, NSR, Noise, Other	[188]

MultiFusionNet	Raw ECG, RRI	80%	-	-	-	CinCDB	AF, NSR, Noise, Other	[177]
CNN, LSTM	LS	82.1%	-	-		CinCDB	AF, NSR, Noise, Other	[201]
CNN, RNN	Raw ECG	82%	-	-	-	CinCDB	AF, NSR, Noise, Other	[190]
DNN	Raw ECG	83.7%	-	-	-	Private	12	[65]
1D-CNN	Raw ECG	-	-	-	91.33%	MITDB	17	[194]
CNN, SVM	spectrograms	-	-	-	97.23%	AFDB,NSRDB, MVFDB, ESTDB	4	[152]
ResNet	Raw ECG	-	-	-	98.6%	MITDB	5	[198]
MCNN	IHR	-	98.22%	98.11%	98.18%	AFDB	AF, Other	[193]

*CUDB—CU Ventricular Tachyarrhythmia Database (CUDB).
*MITDB—MIT-BIH Arrhythmia Database.
*VFDB MIT-BIH malignant ventricular arrhythmia database.
*MSResNet multi-scale decomposition enhanced residual convolutional neural network.
*ESTDB-European ST-T Database [174].

2.6 Research gaps

Based on the aforementioned description, the number of research gaps can be summarised as follows.

1. The ambulatory ECG collection systems for longitudinal arrhythmia collections suffer from sustained patient engagement and compliance [147, 202]. They do not provide adequate feedback to keep patients motivated for longitudinal monitoring. The traditional wired Holter monitors are bulky and restrain the patients' movements [85, 159]. The mobile-based context-aware ECG collection systems (discussed in section 2.3) offer limited context collection, and they too do not address the issue of lack of user engagement. Furthermore, their usability, feasibility, and clinical usefulness of contextualized ECG in improving arrhythmia screening remain unexplored.

2. Like any other patient-generated data, the paper-based diary [142] used for reporting the symptoms/events (e.g., dizziness, palpitation, shortness of breath) encountered during the recording period suffers from recall bias and noncompliance [202, 184] as patients do not always carry it with them. The mismatch in timestamps of the reported symptoms with ECG is challenging during the analysis and limits the diagnostic value of ambulatory Holter monitoring [202].

3. Public benchmark arrhythmia datasets (described in Table 2.1) are collected in a controlled clinical environment with clinical-grade ECG Holter monitors and are therefore relatively clean [55]. In addition, they are manually corrected (e.g., manual correction of R-Peaks) and do not represent the poor signal quality and confounding artifacts expected in patient-operated wearable ECG monitors under free-living conditions. Therefore, they are not adequate for training and evaluating arrhythmia classification models, which are expected to work on wearable ECG under free-living conditions. Most importantly, existing public databases do not provide information about the patient's ambulatory context (e.g., physical conditions, lifestyle), in the absence of which ambulatory ECG analysis remains prone to misclassification [40, 43].

4. The majority of the state-of-the-art DL-based arrhythmia detection algorithms in Table 2.2 are evaluated on the public open-access arrhythmia databases listed in Table 2.1. They tend to give high performance on these datasets. Their generality is a major concern [40, 43]. Usually, applying the arrhythmia detection algorithms trained on these public datasets to patient-operated ECG recordings from free-living ambulatory conditions result in non-trivial FPs detection and displays general performance degradation [64, 135]. These algorithms do not factor in the patients' free-living ambulatory context [40, 43], in the absence of which the noises/artifacts mimicking arrhythmias [116] could easily lead to the wrong diagnosis [45, 40]. The FPR (even if small) in longitudinal monitoring could lead to over-diagnosis and patient anxiety [28, 96]. Furthermore,

in the clinical setting where cardiologists manually verify the arrhythmia detection algorithms' output, the high FPR in longitudinal screening increases cardiologists' workload. Therefore, to enable mobile- and wearable-based longitudinal arrhythmia screening under free-living conditions, the problem of the FPR needs to be addressed.

To address the problem of lack of patient engagement, recall bias in the patient-reported symptoms/events diary, and availability of patient's free-living ambulatory contexts, in this thesis we designed, developed, and evaluated mCardia – a contextualized ECG collection for longitudinal arrhythmia screening [A.2]. To address the limitation of public databases, we designed and developed CACHET-CADB [A.4]. It provides a 259-day contextualized ambulatory electrocardiography arrhythmia dataset from a patient-operated ECG under free-living conditions. To address the problem of FPR under free-living ambulatory conditions, we investigated the patient's ambulatory contexts that induce the non-trivial FPR in a state-of-the-art arrhythmia detection algorithm [A.3]. Thereafter, based on those findings, we proposed *DeepAware* – a new hybrid AF detection model that combined DL with context-aware heuristics. By using context-awareness, *DeepAware* significantly improves the AF under free-living conditions and lowers the FPR [A.5].

CHAPTER 3

Technology for Collecting Contextualized ECG for Arrhythmia Screening under Free-Living Conditions

This chapter addresses RQ 1. The content presented here is from the article [A.2], which details the design and development of a contextualized ECG collection system for arrhythmia screening in patients' free-living ambulatory conditions. The proposed system is named as *mCardia*. The outline of the chapter is as follows. First, section 3.1 describes the motivation behind designing the *mCardia* system. After that, section 3.2 presents the research methodology used in designing the *mCardia* system. The details of its user interface and software architecture and implementation are described in section 3.2.3 and section 3.3, respectively. Finally, section 3.4 summarizes the chapter.

• • •

3.1 Motivation

To investigate the main RQ, first, we need to understand what contextual information about the patient's ambulatory setting is relevant for improving arrhythmia analysis under such free-living conditions. Secondly, as mentioned in chapter 2, the longitudinal patient-operated ECG monitoring for arrhythmia screening suffers from the lack of sustained patient engagement, recall bias in patient self-reported diary data, and poor quality ECG signals [147, 185, 202]. The poor quality ECG profoundly reduces the diagnostic quality during manual and computer-aided automatic analysis [148, 126].

For collecting the contextualised ECG, a few context-aware ECG collection systems have been introduced [164, 100, 109, 123]. However, the context collection in these systems is limited to activities or location. The traditional wired Holter monitors are inadequate for longitudinal screening as they are bulky, and without any feedback to the patient, they remain a black box [85, 159]. To overcome these limitations in longitudinal arrhythmia screening under free-living conditions and collect the relevant ambulatory contexts that can help improve the arrhythmia diagnosis, the task of designing the mCardia system was undertaken. Figure 3.1 shows the schematic diagram of *mCardia*. Through the design of mCardia, we aim to:

- Identify the relevant contextual information that can help in improving arrhythmia diagnosis under free-living ambulatory conditions.
- Build a Holter independent, plug-in based system that integrates with any new ECG Holter and other devices (e.g., blood pressure monitor) without modification in the core functionality.
- Overcome the problem of recall bias in patient-reported event-diary and lack of patient engagement in longitudinal arrhythmia screening.
- Facilitate doctor and patient communication in a chronic care model.

3.2 Research methods

This section describes a brief overview of the research methodology used for designing the mCardia system. For further details of design process and each task, please refer to the article [A.2].

The *mCardia* system design involved a user-centered design (UCD) approach [63] and applied the patient-clinician-designer (PCD) framework [115]. The UCD is an iterative design process used by developers and designers in software product development. It ensures that the product (1) meets the end-users needs, (2) understandable and usable, (3) fulfills the desired task, and (4) provides a positive and enjoyable user experience [132]. The mCardia design was done in collaboration with clinicians and patients affiliated with the cardiology department at Bispebjerg and Frederiksberg Hospitals in Copenhagen. It involved six patients, four clinicians (one Holter nurse

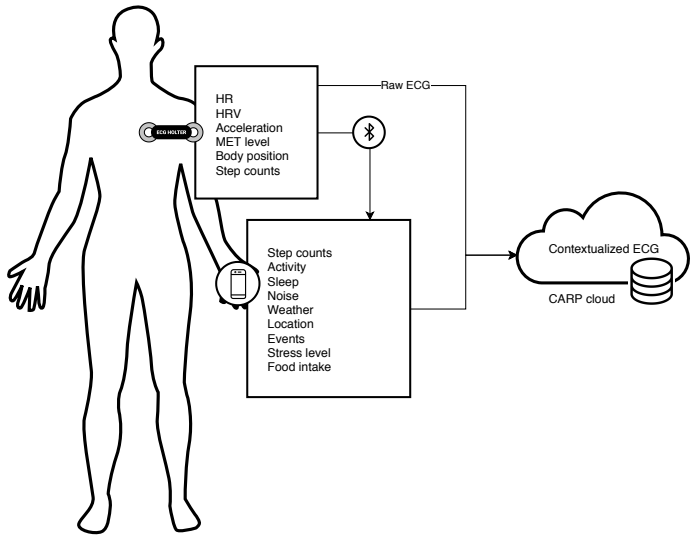


Figure 3.1: Schematic diagram of *mCardia* system. The contextual and physiological data collected from ECG Holter and phone are transmitted from the phone to a cloud-based data management system (image from [A.2]).

and three cardiologists). The timeline of the design process and its main tasks are illustrated in Figure 3.2. The tasks included: identifying the system's context of use, requirements specification, and two iterative design and evaluation steps.

3.2.1 Identifying the context of use

Many design meetings and workshops involving a cardiologist, a nurse, and three patients were conducted. The focus of these meetings and workshops was on designing the following three key aspects of the mCardia system:

- What physiological and contextual data needs to be collected?
- The user experience design of entering the subjective contextual data, data visualization, and app navigation.
- Understanding how the contextual and self-reported data can be used in clinical practice to improve arrhythmia screening.

Alongside the design meetings and workshops, an observational field study was carried out at the Bispebjerg and Frederiksberg hospitals' outpatient arrhythmia clinic to understand the current Holter monitoring process (Figure 2.5). The three-step process of (1) preparing and setting up a patient for Holter monitoring, (2) introducing and mounting Holter on patients, and (3) receiving the Holter and data back from the

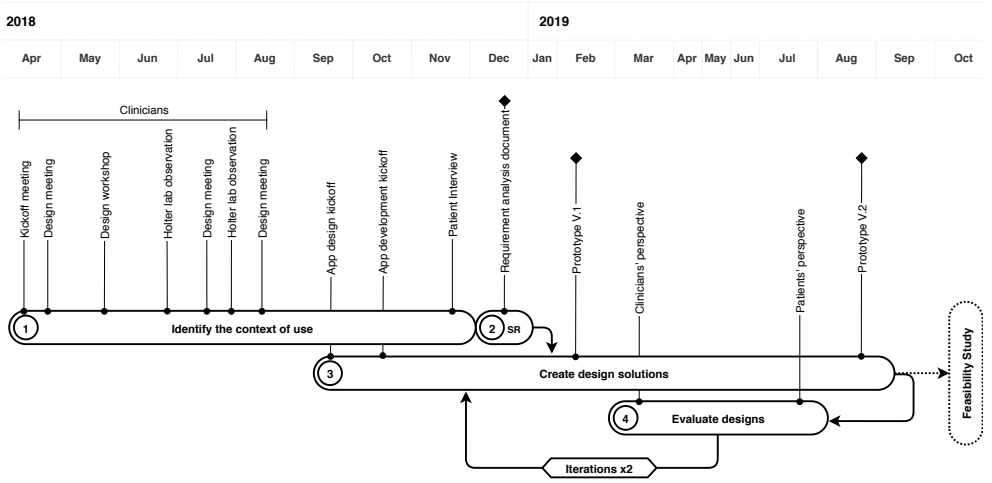


Figure 3.2: Timeline of the UCD process. SR: Specify Requirements (image from [A.2]).

patient were observed from both clinician and patient perspectives. The field study also reviewed current software systems used for processing and analyzing the ECG and patients' self-reported paper-based diaries. Throughout this field observatory study, detailed notes were taken.

Three patients who had prior experience of using traditional Holter monitors were interviewed to understand their requirements and the context of using the system. The open-ended interview inquired about their experience of using the Holter. During these interviews, detailed notes were taken.

3.2.2 Requirements specification

Following the design interviews, workshops, and field study of the traditional Holter process, the requirements specification were finalized as follows:

User engagement in data collection: In terms of user experience and engagement, both patients' and clinicians' were interested in collecting and mapping self-reported symptoms and activities. In the traditional Holter monitoring process, self-reported symptoms and activities are collected using a paper-based diary. During the ECG analysis, clinicians need to map this to the ECG recordings. Clinicians pointed out that the process of manual mapping of symptoms and activity diary with ECG was cumbersome. Often data was not valid as the paper-based event diaries suffered from several flaws, including recall bias, missing and incomplete data. Patients do not carry the paper diary and pen with them at all times. Details were often filled based on recall memory, leading to a mismatch of reported symptoms and their reflection in the ECG. For instance, we observed that P1, who took the Holter test and

participated in our design process, only entered activities such as cycling and running at the end of the day and often forgot to enter details; P2 did not record anything at all. P3 shared that she usually noted down things on her phone and manually added them to the paper diary at the end of the day. They also described that it was often difficult to push the event marker button on the Holter device for registering the events as it was difficult to locate it beneath clothes. Therefore, to improve the user experience and engagement for collecting self-reported symptoms and automatically mapping them with the timeline of the ECG recording, a better method was required. A smartphone-based alternative for event capturing was discussed and brainstormed during the workshop. All the patients unanimously agreed that recording the symptoms and activities in a smartphone app would be much easier than a paper diary, as they always carry their smartphones with them.

Collecting contextual information relevant for arrhythmia screening: From the literature review of common arrhythmia triggering factors [66, 62] and clinicians' interview list of relevant contextual information was prepared. It included physical activities (e.g., walking, running), body movements/positions of the patient (standing, sitting, laying down, turning side in the bed), self-reported symptoms, symptom duration, patients' activity during the symptom, location, step counts, stress level, sleep, food intake (e.g., light, heavy), and surrounding environment (e.g., temperature, noise level). The process of collecting this broad set of contextual information would involve implementing the ecologically momentary assessment (EMA) approach and automatic context collection from the on-board sensors of the ECG device and the smartphone.

System feedback: To ensure quality data collection, keeping patients engaged in the data collection process is essential and requires providing patients some feedback during the ECG collection period. The clinicians were against the idea of providing visual feedback on physiological data to the patient. They argued that given the complex nature of the ECG data, it might not be meaningful for the patient, and it could cause unnecessary concerns and anxiety. On the other hand, all the patients were very interested in seeing their ECG data and welcomed the idea of getting any feedback that the system can provide about their heart condition. Employing the PCD approach, we found a trade-off among these conflicting requirements. We finalized that the mCardia should not provide any system-generated feedback that might cause anxiety for the patients (such as automatic detection of AF), and at the same time, provide an overall visualization of selected data items to keep them engaged.

3.2.3 Iterative system design and mCardia user interface

Based on the aforementioned requirements specification, a Minimum Viable Product (MVP) was outlined. The Movisens 'EcgMove4' ECG device was chosen for the MVP due to its availability, data collection features, and its open application programmer interface (API). The system was designed and implemented in two major iterations.

For the details of each iteration of system design, we direct the reader to article A.2.

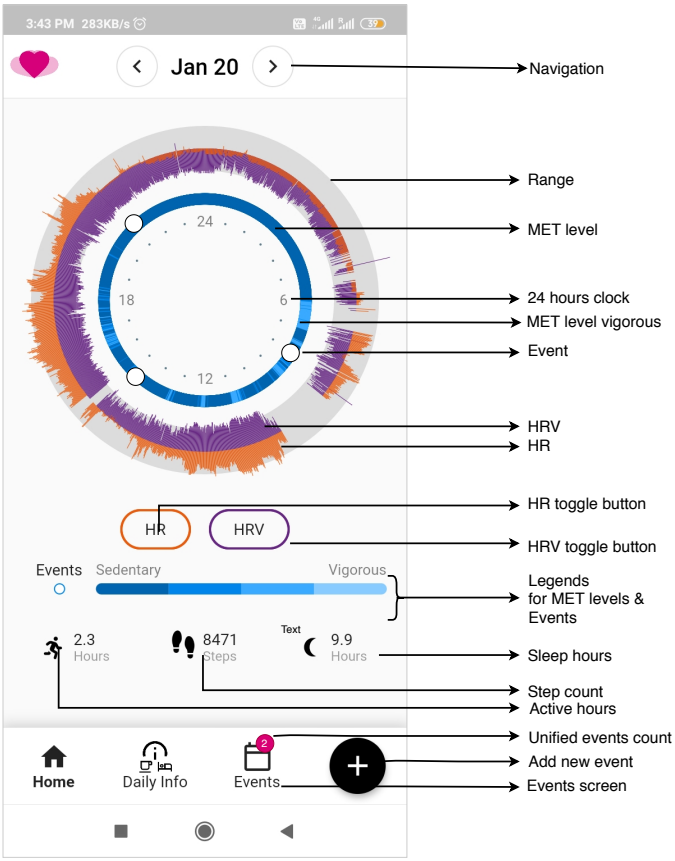
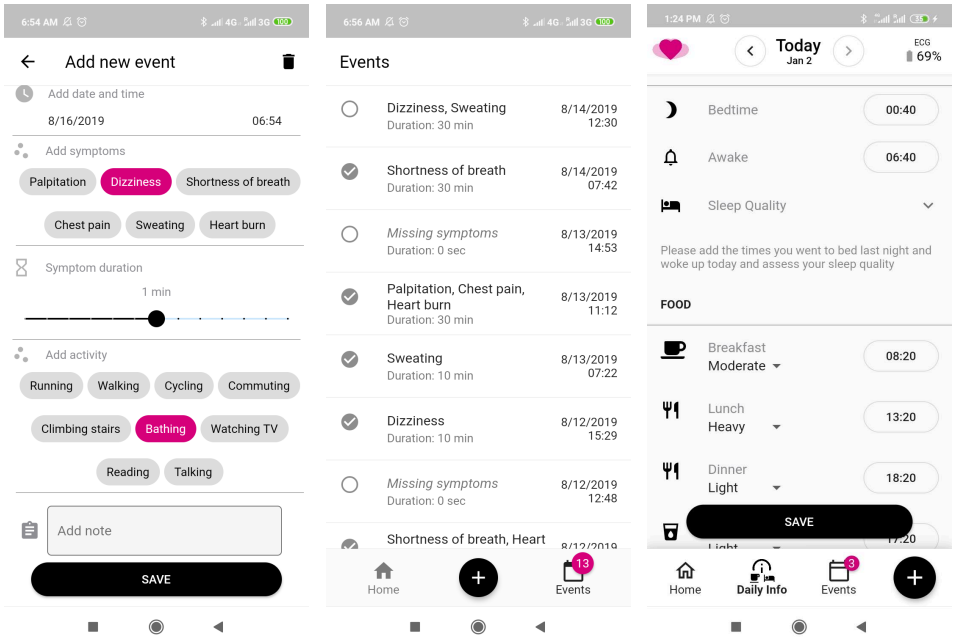


Figure 3.3: Home screen of *mCardia* (final design). The 'wheel' is the main UI element which presents the detailed recordings of HR, HRV, and MET-level in a 24-hour clock (images from [A.2]).

mCardia's Final User Interface: Figure 3.3 and 3.4 show the screenshots of the final mCardia mobile application. A brief overview of the different menus is as follows. Figure 3.3 is the home screen, and the 'wheel' is its main component which presents the detailed recordings of heart rate (HR), heart rate variability (HRV), the user entered unusual events, and Metabolic Equivalent for Task (MET)-level in a 24-hour clock. The inner and outer range of the circle is between 40 to 100, starting from the middle. With the black plus sign, the user can create a new event (Figure 3.4a), and that will appear on the inner 24-hour clock. Figure 3.4b shows all created events and their status as the "filled/unfilled" or "partially filled" events. Figure 3.4c, on the other hand, shows the screen for collecting daily information such as self-assessed

stress level, sleep quality, sleep timing, and food intakes, including meal timings and types (light, moderate or heavy).



(a) (b) (c)

Figure 3.4: The user interface design of the *mCardia* mobile app. (a) Event details, (b) List of events filled or created by the patient, (c) Daily Info (image from [A.2]).

3.3 Architecture and implementation

To choose a proper framework for implementing mCardia, we searched the existing generic mobile and wearable sensing frameworks that allow building high-frequency contextualized data collection applications. Specifically, we were interested in a framework that (i) supports cross-platform (both Android and iOS) and preferably written in a single programming language, (ii) is up-to-date and maintained with proper plug-in/API/architecture documentation, (iii) can handle high-frequency data and provide seamless support of data synchronization with the cloud, (vi) supports a plug-in architecture for adding any new ECG Holter as plug-and-play. A systematic review of the different mobile and wearable sensing frameworks was carried out – further details which are presented in the article [A.1].

Excluding a few (e.g., Aware [50]), most frameworks in the review were either not up to date or lacked documentation. Even up-to-date or documented once did not meet our requirement of supporting multi-platform with a single programming language. Therefore, we decided to use our newly developed frameworks, CARP Mobile Sensing (CAMS) [18] and the Research Package [73] for building mCardia. They are built on Google's cross-platform development toolkit called as Flutter and written in Dart programming language. CAMS offers support for building cross-platform and extensible mobile sensing apps and can seamlessly handle high frequency data collection, data anonymization, data upload, and power (battery) optimization. The Research Package framework supports the *GDPR* complied informed consent handling and delivering ecological momentary assessment (EMA) questionnaires.

Figure 3.5 shows the software architecture of mCardia. Each of these frameworks comprises several sub-components (marked green in Figure 3.5). The mCardia's data sampling is configured as a 'Study' script in CAMS. The sampling packages are registered with the study controllers, which encapsulated access to the operating system (OS) sensors and are responsible for handling the data sampling. Any new plug-in can be integrated with CAMS via registering its sampling package. For integrating the ECG Holter (Movisens EcgMove4 in this prototype), we implemented *MovisensSampling Package* [97]. It usages a `Movisens` Flutter plug-in [98] to access the native Movisens application programmer interface (API) via Bluetooth low energy (BTLE) (marked in purple in Figure 3.5). Please note that due to its modular and plug-in based architecture, any new ECG Holter can easily be integrated with *mCardia* simply by registering its sampling package without making any changes to the application. The implemented *mCardia* system was published on the Google Play store[1] in December 2019.

For further implementation details of individual components of mCardia, we direct the reader to the article [A.2].

3.4 Summary

In this chapter, a brief overview of user research, design, development, and technical aspects of the mCardia system were presented. The mCardia is designed for collecting context-aware ECG for arrhythmia screening in patients' free-living conditions. The plug-in based architecture of mCardia will allow it to be integrated with any new ECG Holter without making any changes in the core application. Compared to the traditional ECG monitoring Holters and existing context-aware ECG collection systems, mCardia facilitates collecting a wide range of contextual data. The relevant contextual information for improving ambulatory arrhythmia diagnosis includes patient's activities, movement acceleration, body position (sleeping left/right/supine), self-assessed sleep quality, stress level, self-reported unusual events (e.g., dizziness, palpitation), and food intakes. These contextual data are synchronized with the raw

[1] `https://play.google.com/store/apps/details?id=com.cachet.reafelapp`

ECG timestamp. The phone-based event diary will solve the problem of recall bias and non-synchronization of patient-reported symptoms/event diary with the ECG, which significantly limited the diagnostic value of collected ECG in traditional Holders [202]. To keep patients engaged and motivated for longitudinal use, it provides patients a contextual overview of HR, HRV, activeness levels, and unusual symptoms/events. In a chronic care model, it can also enhance doctor and patient communication.

In the next chapter, we will evaluate the usability and clinical feasibility of the *mCardia* system through a field deployment study.

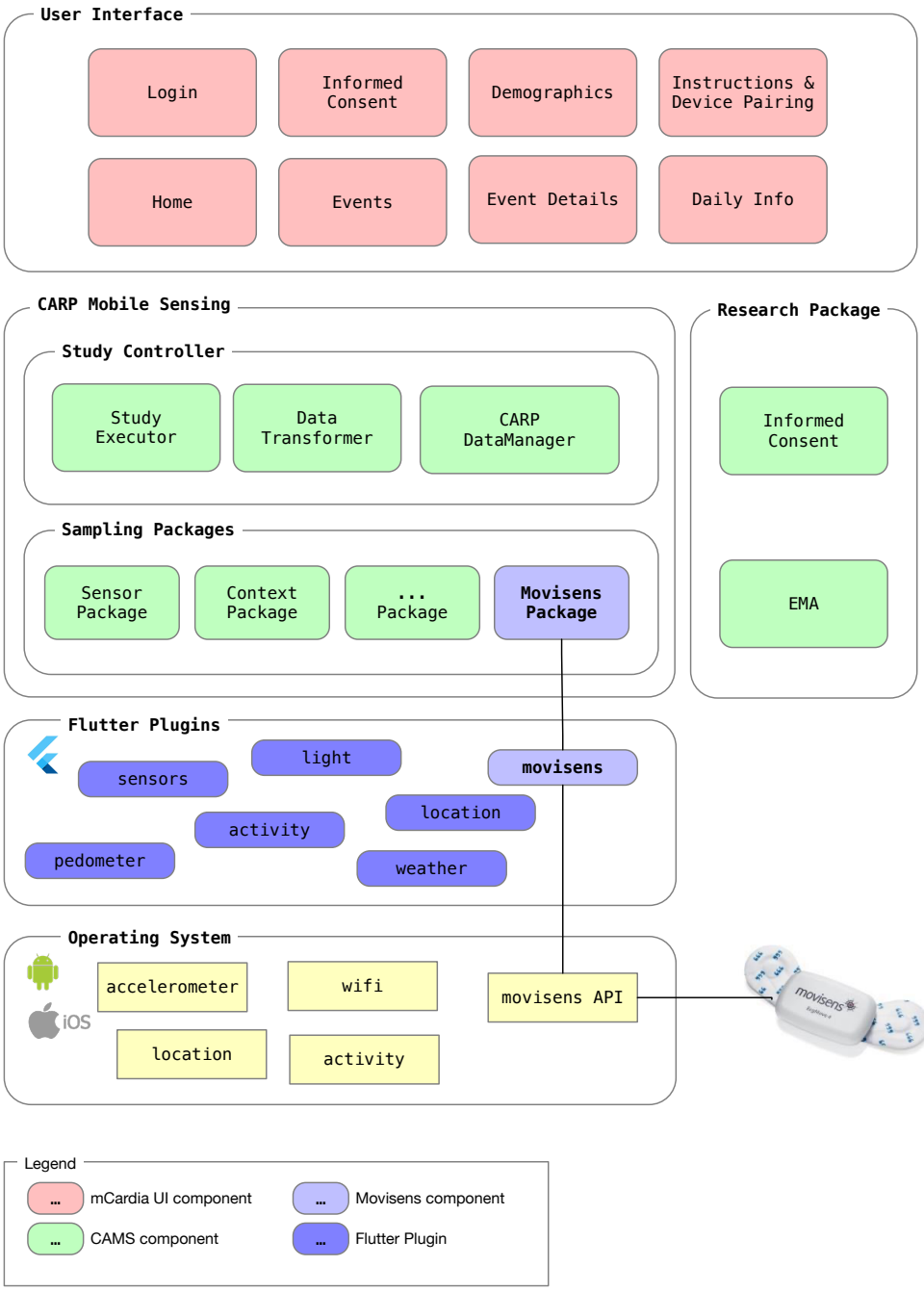


Figure 3.5: The architecture of the mCardia system. The green components are part of CARP Mobile Sensing (CAMS) framework, and the blue components are its associated Flutter plugins. Movisens is a single-channel ECG Holter used in the current prototype of the mCardia system (image from [A.2]).

CHAPTER 4

Usability and Clinical Feasibility Study of *mCardia*

The content of this chapter is based on *mCardia*'s feasibility study reported in article [A.2]. It addresses RQ 1 by assessing the usability and clinical feasibility of the *mCardia* system in collecting contextualized ECG under ambulatory free-living conditions through field study. The outline of the chapter is as follows. Section 4.1 describes the design of the usability and clinical feasibility study. Thereafter section 4.2 presents the study results. Finally, section 4.3 summarises the chapter and lists the lessons learned during the feasibility study.

4.1 Study design

It is argued that, in the early stage of novel health technology, the "how and why of a system used by its target users" needs to be evaluated [95]. Therefore, following the design research's best practices, a single-arm feasibility study was planned to assess *mCardia*'s usability and feasibility under free-living conditions. In a patient's natural setting, sustained engagement and collecting quality ECG data is a challenge; therefore, the following aspects remained the main focus of the mCardia's feasibility study.

1. The usability evaluation–which included perceived user engagement, usefulness, and usability of *mCardia* in longitudinal ECG data collection.

2. Technical robustness and feasibility of *mCardia* in collecting and managing the contextual and physiological data.

3. Clinical usefulness of collected contextual data in the arrhythmia screening process.

The feasibility study was conducted in India and Denmark. The research protocol for the *mCardia*'s feasibility study was reviewed by Danish Research Ethical Committee and exempted from ethical approval as the study did not involve any clinical intervention or treatment (File # H-19071015). Similarly, in India, the study approval was obtained from the institutional ethical committee of the Mahatma Gandhi University of Medical Sciences & Technology (MGUMST), Jaipur. The patients signed the informed consent and allowed their anonymized data to be used for any research and analysis purposes.

4.1.1 Patients recruitment

Table 4.1 lists the inclusion and exclusion criteria used for assessing the patients' eligibility for the study. Patients were recruited during their outpatient arrhythmia clinic visit via a general announcement to participate in the study. The patients were informed that study data would not be used for their ongoing treatment or diagnosis.

Inclusion	Exclusion
Previously undiagnosed, however, were at high risk or suspected of having arrhythmias	Age below 18
Patients already diagnosed with AF but interested in tracking AF symptoms	Hospitalized/bedridden or in critical health conditions
Comfortable in using smartphone apps and wearables or have a caretaker/family member who can help them in using *mCardia*	
Willing to use *mCardia* for a minimum of two week	

Table 4.1: An overview of the inclusion and exclusion criteria for the study.

4.1.2 Study procedure

A detailed description of the feasibility study procedure is provided in the article [A.2]. Figure 4.1 gives an overview of the study procedure. Figure 4.2 shows patients using *mCardia* system in their free-living conditions.

4.2 Results

For a thorough exposition of the feasibility study's results and reflection, we direct the reader to article [A.2]. However, the key findings are summarised below.

A total of 33 patients were recruited, of which 9 dropped out and could not complete the minimum 2-week study period. The study results are based on the

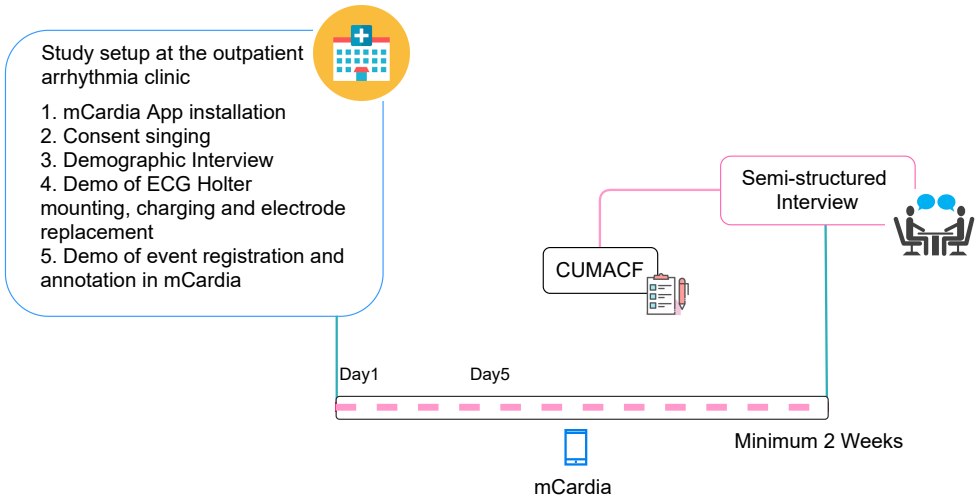


Figure 4.1: An outline of the *mCardia*'s feasibility study procedure. Each pink box denotes a single day in the study duration. CUMACF: CACHET Unified Methodology for Assessment of Clinical Feasibility [16, 15]. The CUMACF questionnaires are available in Appendix B.

Number of patients	24
Gender – Female / Male	8 / 16
Age – Mean (SD)	58.79 (10.11)
Prior ECG Holter experience	10
Assisted by caregivers	9

Table 4.2: Demographics of the patients (Table from [A.2]).

quantitative and qualitative data analysis of the remaining 24 patients. Table 4.2 shows the demographics of the patients.

4.2.1 User experience

Perceived usefulness and usability: Figure 4.3 shows the result of *mCardia*'s usability and perceived usefulness as obtained through post-study CUMACF (CACHET Unified Methodology for Assessment of Clinical Feasibility) questionnaires. The overall response for *mCardia*'s interface and usability was positive from 96% of the patients, and it was perceived as unobtrusive and non-interfering in their everyday activity (Q1). Approximately 95% of the patients answered that keeping track of daily activity and unusual symptoms (as done in *mCardia*) could help in giving a bet-

(a) (b)

(c) (d)

Figure 4.2: Patients using *mCardia* system in their free-living conditions.

ter understanding of their symptoms and overall health (Q5). Nearly 78% opined that *mCardia* could help them better communicate with their doctor (Q3). The prospec-

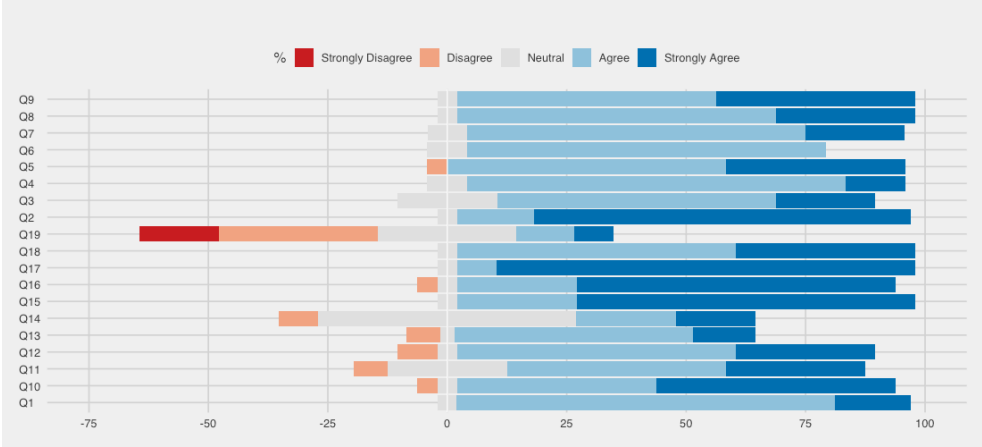


Figure 4.3: The CUMACF [16] based perceived usability and usefulness scores of *mCardia* system. The questionnaires (Q1–Q19) are provided in the Appendix B (image from [A.2]).

tive of reducing the recall bias (Q4) in reporting events/symptoms in home-based longitudinal ECG monitoring was especially liked by the patients who previously underwent the home-based Holter monitoring with traditional wired Holters. During the post-study interview, one of such patients commented that:

> *"The big wired Holter monitor was just a black box for me, and it was not very comfortable to sleep or work while wearing it. Also, I was not strict about keeping the symptoms diary, as I would not keep the diary and a pen with me at all times."* [P23]

Engagement as time spent on *mCardia*: Although *mCardia* was designed for brief sporadic use, the screen time is a good indication of user engagement with mobile apps. We looked at the amount of time spent by the patients and found that, on average, patients spent daily 21 minutes actively interacting with *mCardia*. The daily information such as food intake, sleep, stress was mostly entered at once in the evening. The interactions were more common in the morning (to get an overview of night heart rate) and after any physical activity/exercise.

Unusual events recording and phone-based context annotation Figure 4.4 shows the number of unusual events annotated or registered (and deleted) over 300 days of the study period. Out of 235 total registered events, nearly 60% were partially or fully annotated–the rest 40% remained either un-annotated or were deleted as they were registered due to accidental tapping of ECG Holter. The events registered (and deleted) due to accidental taps on the ECG Holter were higher in the initial days. However, after that, it significantly dropped as patients became accustomed to *mCardia*.

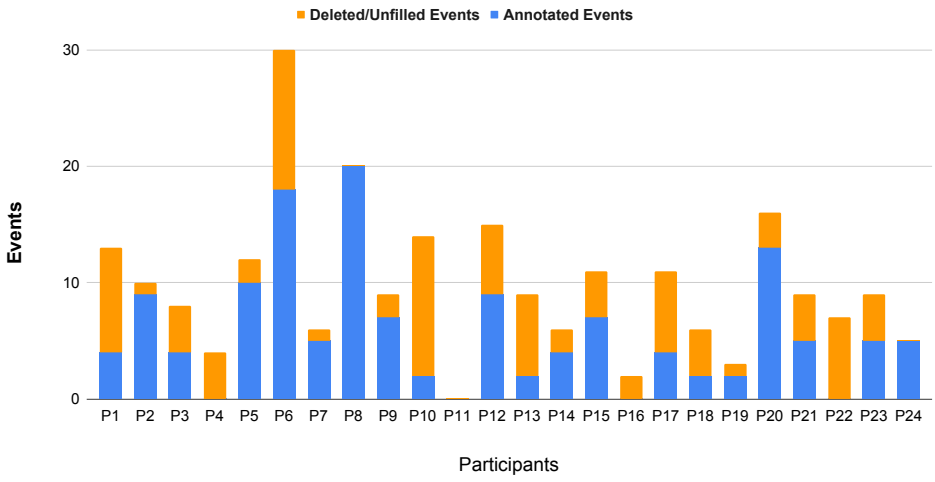


Figure 4.4: Per participant number of annotated or deleted/unfilled events (image data from [A.2]).

4.2.2 Performance of data collection

Data quality: Over 8064 hours long contextualized ECG was collected, of which 89% was found suitable for arrhythmia analysis. The percentage of unusable/noisy data was more among the patients assisted by the family/caretaker during the recording period. In the post-study interview, we found that when electrodes become nonadhesive in such patients, they would not realize it as they themselves were not viewing the *mCardia*. Although *mCardia* shows a gap in data when electrodes are nonadhesive, caretakers would not realize it and ask the patient to replace ECG electrodes until they opened the *mCardia* on the patient's phone. As a caretaker explained:

> "I would usually check and change the electrodes only when I found gaps in the HR or HRV data in the *mCardia* app's circular wheel." [P3]

Yield: The yield of a system is defined as:

> "The fraction of the expected samples to the actual number of samples collected by the system" [108].

It has been argued that for determining engagement and user compliance, yield can serve as a proxy [108]. Figure 4.5 shows the yield of various physiological and contextual data types collected in *mCardia*.

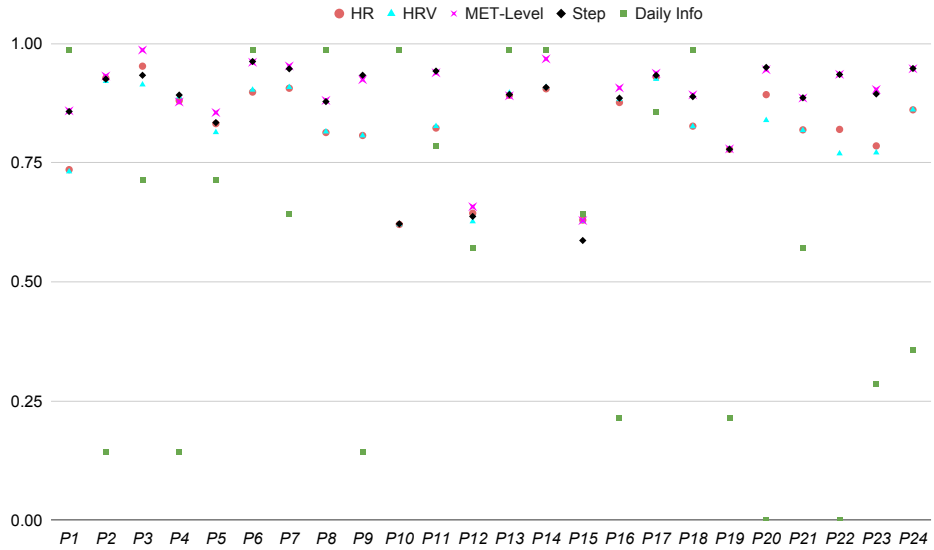


Figure 4.5: Data collection yield of various physiological and contextual data (image data from [A.2]).

4.2.3 Qualitative analysis of post-study interview

An inductive approach [175] was used for analyzing the post-study interview and the patient's response to CUMACF questionnaires. The following themes were in focus; (i) issues encountered, (ii) task suitability, and (iii) perceived usefulness.

Issues encountered: Table 4.3 lists the issues identified and their relative distribution among study patients. Accidental tapping on the Holter device resulting in false events was one of the main issues. As recalled by a patient:

> *"For the first two days, when I pushed the ECG device hard in order to fit it into the charging tray, it added some events. When I saw these empty events later in the app, I was confused as I didn't recall tapping on the device."* [P10]

Task suitability: Task suitability was focused on understanding the suitability of phone-based events and other manual contextual data collections. The phone-based annotation of unusual symptoms and events was most appreciated by patients (N=10) who previously have had the experience of traditional home-based Holter monitoring and writing paper-based event diaries. As described by P20:

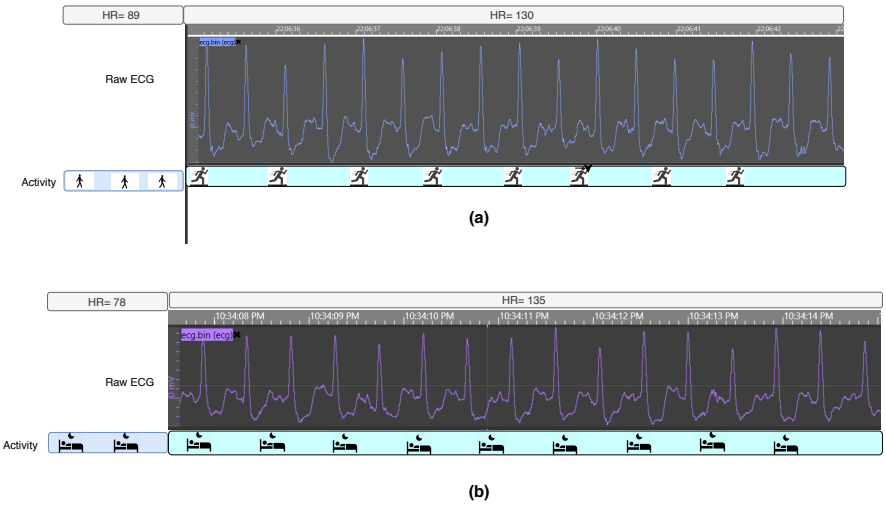


Figure 4.6: Clinical case #1 demonstrating the usefulness of context information in ECG interpretation. Although both ECG snippets from the same participant have HR>130, the context information (sleeping or running) helps in distinguishing that (b) is a case of Supraventricular Tachycardia (SVT), whereas (a) is normal (image from [A.2]).

Issue Identified	**patient(%)**
Electrodes becoming non-adhesive	26%
Skin irritability due to continuous use of wet ECG electrodes	20%
Accidental tapping on ECG device causing false event registration	38%
ECG device battery discharged	20%

Table 4.3: Issues identified and their relative distribution (Table from [A.2]).

> *"It is easier to note the symptoms on the mobile phone, since I carry it all the time. I do not have to remember and write it down in my [paper-based] event diary–especially, if I have to do it for many days or months."* [P20]

Perceived Usefulness:

The patients and their caretakers described that they considered *mCardia* could be helpful in two ways; (i) in better communicating their symptoms to clinician, and (ii) to keep them aware of any unusual symptoms/events and the context in which they appear. For example, as explained by P2:

> *"I think this clock overview is nice. I can see how my heartbeat changes*

when I am doing different activities. On two Fridays–when I had been playing basketball–I felt palpitations during the night and registered these events. It makes it easy to show them to the cardiologist and ask what happened at that time, and if it is related to playing sports." [P2]

4.2.4 Clinical usefulness of the contextual data

Figure 4.6 and 4.7 demonstrate two clinical cases where contextual information helped cardiologist in making better interpretation of ambulatory ECG during the manual analysis.

Clinical case #1: The resting HR above 100 BPM in adults is defined as Tachycardia [13]. As shown in Figure 4.6, although both ECG snippets from the same participant have HR>130, the contextual information (sleeping or running) aids in identifying that (b) is a case of Supraventricular Tachycardia (SVT), whereas (a) is normal. This case illustrates that depending on the patient's ambulatory context, ECG segments and HR can have a different interpretation, and evaluation in isolation could potentially cause misdiagnosis.

Clinical case #2: Clinical case in Figure 4.7 shows the advantage of context-awareness in medication prescription for the arrhythmia patient. The patient was evaluated for annoying palpitations and had known permanent AF. Figure 4.7 (a) shows the subject tapped on the device and registered an event of palpitations during sleep that lasted 30 minutes. It also reveals cases of mild rate changes, whereas the accelerometer data below shows that the patient is just turning in bed. At this point, his doctor may hesitate to increase rate-lowering medications as the patient also reported dizziness while changing from laying down to a standing position. In this context, Figure 4.7 (b) aids in determining an adequate medicine adjustment as it is reassuring that there is a sufficient chronotropic response when switching from laying in bed to walking. It suggests a problem with orthostatic blood pressure changes. Moreover, there are no cases of severe low rates. Therefore, the choice could be made for a rate-lowering medicine without a blood-pressure-lowering effect. This analysis demonstrates that besides helping in arrhythmia diagnosis, ambulatory contextual information can also help in determining correct medication prescription.

4.3 Summary

The outcome of the feasibility study was as expected. It has achieved high perceived usefulness and usability scores. The *mCardia* system was engaging for the patients and capable of collecting quality ECG data (as reflected in 89% of the usable data for arrhythmia analysis) under free-living conditions. The practice of reporting and annotating the symptomatic events on the phone rather than a paper-based diary was especially liked by those patients who earlier had undergone home-based ambulatory Holter monitoring using the traditional Holters. In partial answer to main **RQ**, the

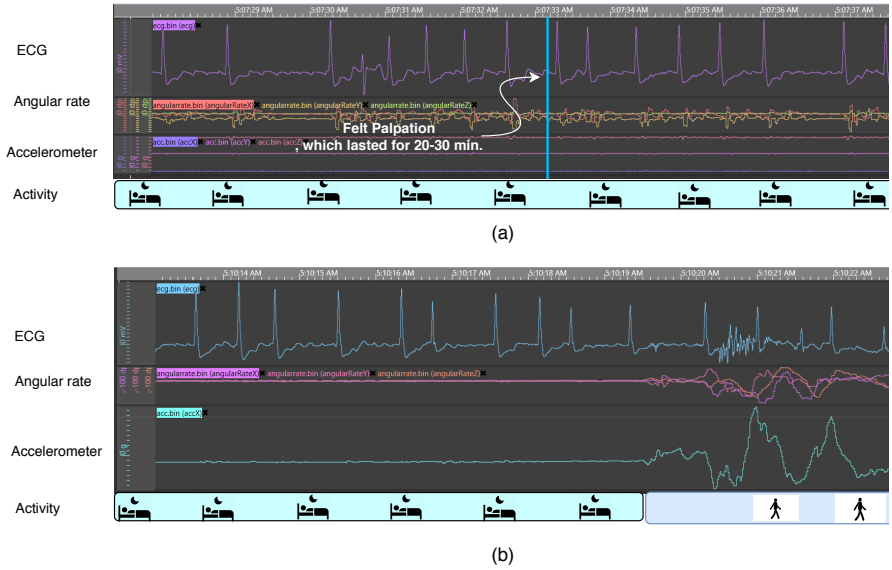


Figure 4.7: Clinical case #2 demonstrating the usefulness of contextual information in medication prescription. (a) Subject tapped on device and registered an event. (b) Subject gets up from the bed and starts walking (image from [A.2]).

two clinical case studies of manual arrhythmia analysis reflected how contextual information helped in making a better assessment of the ECG from free-living ambulatory conditions.

From the perspective of further answering the main RQ and RQ 2 (in particular) via the *mCardia* system development and its feasibility study, we were also interested in collecting a contextualized ECG arrhythmia dataset. In the next chapter, this dataset will be used to answer RQ 2. We will investigate the temporal and contextual distribution of FPR in a state-of-the-art DL-based AF detection algorithm when it is applied to patient operated ECG from free-living ambulatory conditions.

CHAPTER 5

Impact of Ambulatory Contexts on False-Positive Rate: An Investigation

The content of this chapter is from the article [A.3]. It addresses RQ 2 by investigating the impact of ambulatory contexts on FPR in a state-of-the-art AF detection algorithm when applied to ECG data collected under free-living conditions. The outline of this chapter is as follows. Section 5.1 describes the motivation behind the investigation. Thereafter, section 5.2 provides an overview of the experimental setup. Finally, the results and summary of the findings are described in section 5.3, and section 5.4, respectively.

5.1 Motivation

As mentioned in the introduction, to facilitate the automatic analysis of longitudinal ECG, many computer-aided arrhythmia detection algorithms have been developed. These algorithms have evolved from the traditional feature engineering-based approach (described in chapter 2) to the most recent machine learning techniques such as deep learning. The FE-based approach required manual handcraft feature extraction by a domain expert and is less adaptive to ECG morphology outside of manually extracted features. The DL-based techniques, on the other hand, can enable **end-to-end classification** without manual feature extractions or domain expert's intervention. For this reason, in recent years, the use of DL for AF and other types of arrhythmia detection have gained momentum and have achieved a high level of performance [10, 141, 143, 48, 181, 65, 191, 111, 193].

Despite improvements in the DL algorithms' AF detection capabilities, bringing

them into widespread adoption under free-living conditions remains challenging [40]. The majority of these algorithms are trained and evaluated on public datasets (described in Table 2.1), which contain high-quality ECG collected under controlled clinical environments. These algorithms show high performance on these public datasets. However, when exposed to patient-operated ECG collected under free-living ambulatory conditions, their performance tends to deteriorate and bring on a significant increase in FPs detection [55, 64, 46]. In longitudinal arrhythmia screening, the false positive rate (even as low as 1%) can cause over-diagnosis and patient anxiety [28, 96], particularly in low AF burden population.

For performance improvement, the researchers so far have mainly focused on ECG signal characteristics. However, as pointed out in [43, 40], without understanding the patients' context in which the ECG was collected, ambulatory arrhythmia analysis still remains prone to misclassification. Although it is known that patient's ambulatory contexts add confounding motion artifacts and cause high FPs in arrhythmia detection algorithms [46], it remains unexplored whether there exist any specific correlations between the patient's ambulatory contexts and FP occurrences in a state-of-the-art AF detection algorithm. For instance, there might be the case that a particular type of activity, body position, time, place, or food intake is inducing more FPs in an arrhythmia detection algorithm. A preliminary study in this direction was reported by Noh et al. [130]; they observed that a particular walking pattern (walking on slop) was inducing more FPs in a heartbeat detection algorithm compared to walking on a flat surface or sitting. In our experiment, we want to investigate it more broadly and examine the impact of ambulatory contexts on FPR in a state-of-the-art AF detection algorithm under free-living ambulatory conditions. If specific ambulatory contexts are found to be inducing more FPs, then such information can be used to dynamically fine-tune the algorithm's sensitivity and specificity on those FP-prone contexts. Therefore, understanding the temporal and contextual distributions of FPs in free-living condition can:

- Help in designing DL algorithms that incorporate context-induced AF mimicking artifacts into their design.
- Make end-to-end FPR classification algorithm more transparent, which otherwise remains a black box.
- Identify the sources of the algorithm's mistakes and reduce the FPR by context-specific dynamic adjustment of sensitivity and specificity under free-living ambulatory conditions.

5.2 Research method

A brief description of the research methods is provided here; for details, we direct the reader to the article [A.3]. Figure 5.1 shows the procedure used to investigate the

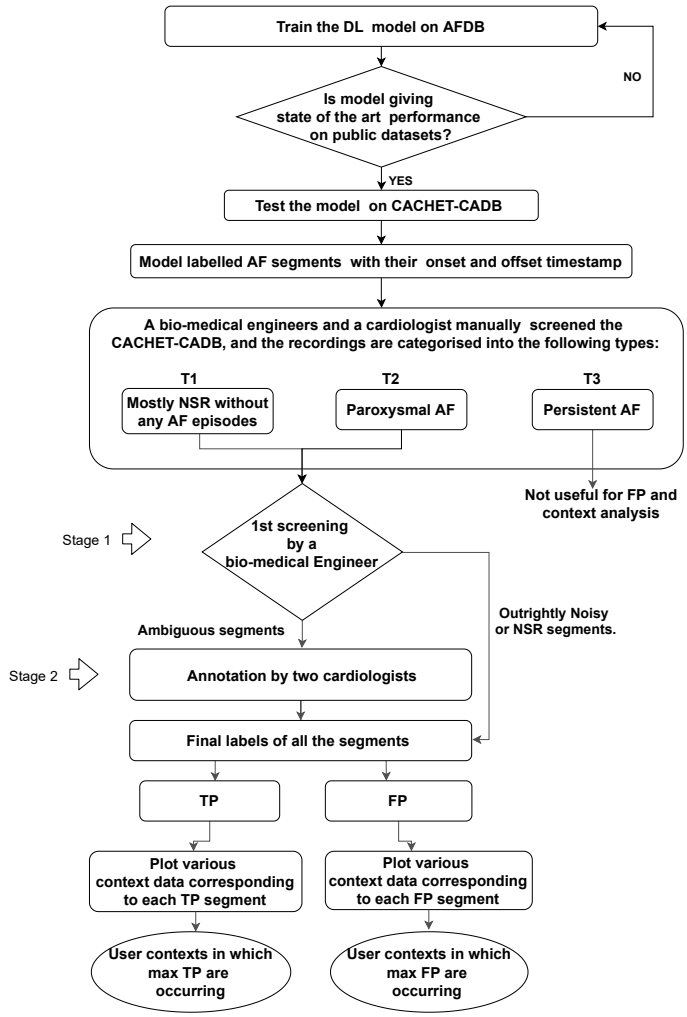


Figure 5.1: Experimental workflow for investigating the influence of ambulatory contexts on FPR. CACHET-CADB is the contextualized ECG dataset collected during the mCardia's clinical feasibility study (image from [A.3]).

contextual and temporal distributions of FPs in a DL-based AF detection algorithm when applied to patient-operated ambulatory ECG from free-living conditions.

- **Step 1:** An end-to-end DL algorithm for AF detection was trained on a public dataset (AFDB), and we ensured that it had a state-of-the-art performance on other publicly available ECG datasets (MITDB and NSRDB). Figure 5.2 shows

the architecture of that DL algorithm. It is a combination of CNN and RNN and trained on RR-interval (RRI) features for the single-channel ECG.

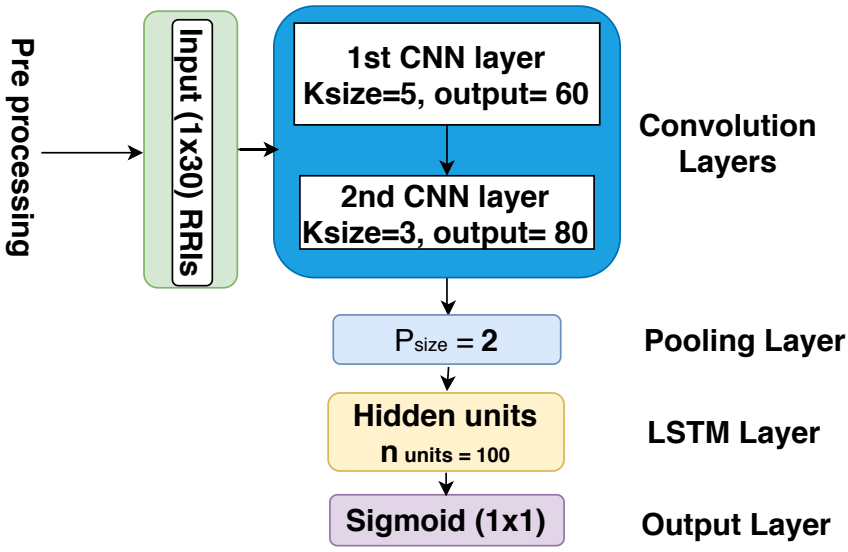


Figure 5.2: The network consisted of 2 convolutional layers, followed by a pooling layer. Features extracted in the convolution are fed into the LSTM layer, which consists of 100 hidden units. Finally, the sigmoid layer gives the binary output of the entire 30 RRIs long window. The network takes 30 RRIs (calculated from raw ECG) data as input and outputs a binary classification of AF or non-AF class (image from [A.3]).

- **Step 2:** The trained algorithm from the previous step, which has state-of-the-art performance on three public datasets, is then tested on a contextualized ECG dataset collected from AF patients under free-living ambulatory conditions. Please note that the contextualized ECG dataset used here is part of data collected during the clinical feasibility study of *mCardia* in chapter 4. The dataset is named CACHET Contextualised Arrhythmia Database (CACHET-CADB). The AF onset and offset segments detected by the algorithms were mapped back from the RRI sequences to the raw ECG timestamps (as depicted in Figure 5.3) and stored in CSV files.
- **Step 3:** In this step, a biomedical engineer and two independent cardiologists manually examined and annotated the AF onset and offset detected by the DL algorithm. A mobile application explicitly designed for ECG annotation was used by the cardiologists. This step produced the ground truth, and each segment were labeled as TP or FP.

- **Step 4:** Finally, after obtaining the ground truth in the previous step, the FP and TP detected by the algorithm were plotted (see Figure 5.5) against the patient's ambulatory contexts (e.g., activities, body positions) for examining their co-relations.

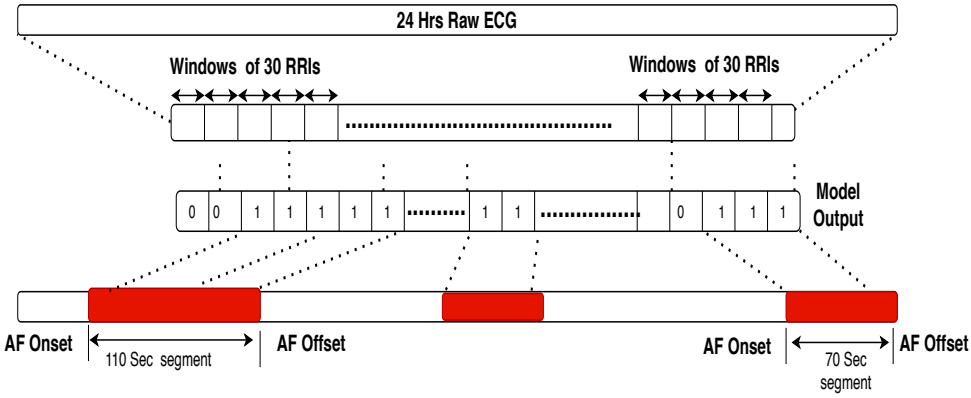


Figure 5.3: Mapping DL-based AF detection algorithm's output to raw ECG timestamp for 24 hours (image from [A.3]).

5.3 Results

In this section, we summarise some of the important results from the above experiment. For details, we direct the reader to article [A.3].

Algorithm's performance on public datasets: The performance of the algorithm on three public datasets from PhysioNet [59] is listed in Table 5.1. On training dataset AFDB, in 5-fold cross-validation, it achieved an accuracy of 97.04% and FPR of 1.7%. Both MITDB and NSRDB are unseen databases for the algorithm and were used only for evaluating the algorithm's generalization. As expected, the FPR on these previously unseen databases was much higher (13.06% and 5.56%, respectively). The NSRDB is an ambulatory ECG dataset from the healthy subjects without any AF, but it still received 5.56% FPR. However, please note that this performance is still at par with other state-of-the-art DL algorithms on these public datasets.

Table 5.1: Performance of the used algorithm on public datasets and its comparison with other state-of-the-art algorithms. Ch: number of ECG channels (Table from [A.3]).

Algorithms	**Ch**	**AFDB**				**MITDB**				**NSRDB**			
		Se (%)	Sp (%)	Acc (%)	FPR (%)	Se (%)	Sp (%)	Acc (%)	FPR (%)	Se (%)	Sp (%)	Acc (%)	FPR (%)
[189]	1	98.79	97.87	98.63	-	-	-	-	-	-	-	-	-
[105]	1	97.4	97.2	97.3	-	-	-	-	-	-	-	-	-
[183]	2	94.28	94.91	94.59	-	-	-	-	-	-	-	-	-
[193]	2	98.22	98.11	98.18	-	-	-	-	-	-	-	-	-
[35]	2	99.93	97.03	96.59	-	-	-	-	-	-	-	-	-
[10]	2	98.17	96.29	97.10	3.71	98.96	86.04	87.40	13.96	-	95.01	-	4.99
This work	1	96.06	98.29	97.04	1.7	96.87	86.94	87.98	13.06	-	94.44	-	5.56

Analyzing false positives in the contextualised ECG data under free living conditions: From the CACHET-CADB, a total of 215 days single-channel ECG data of 21 subjects were analyzed using the trained DL algorithm. Four subjects were found to remain in persistent AF during the manual analysis and were excluded from further analysis (see Figure 5.1). Table 5.2 lists the number of AF episodes labeled by the DL algorithm and the length of these episodes for each patient. Figure 5.4 shows the annotation process summary after manual examination of the algorithm's output.

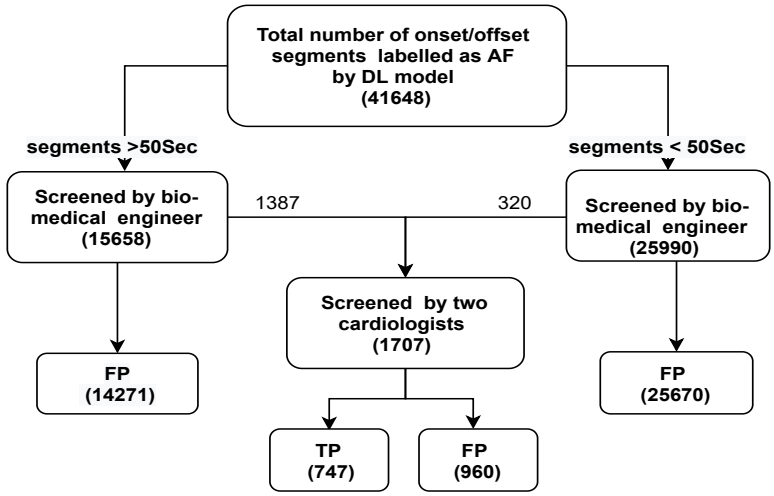


Figure 5.4: Annotations summary after manual examination (image from [A.3]).

Table 5.2 reveals that nearly 62% of the total AF segments detected by the algorithm are of a length less than 50-seconds, and based on annotation results in Figure 5.4 it is apparent that 99.9% of them are FP. Visual inspection of the records revealed that these short segments ($\leq$ 50-seconds) are mostly associated with a change in activity, body position (specially during night), or sudden movement acceleration.

Figure 5.5 presents the AF segments detected by the DL algorithm, the ground truth (i.e., the true labels obtained after manual annotation by the cardiologists), and their correlations with the patient's contexts (activities, body position, and movement acceleration) for 24 hours of ECG. It can be observed that whenever there is a sudden peak in the movement acceleration, and if the segment gets classified as AF, it is mostly FP if it is $\leq$ 50-seconds. The true positive segments were observed in the morning and late evening hours. Also, more FPR was observed on female subjects than male. It could be attributed to (1) the male/female data asymmetry in the training dataset or/and (2) the relatively more confounding motion artifacts in females due to breast movements in a chest-mounted ECG Holter.

Design recommendation: This investigation's finding indicate the influence of the patient's ambulatory contexts on the the state-of-the-art AF detection algorithm's

FPR. Since 99.9% of the short segments (of length $\leq$ 50-seconds) are around change in activity, body position change, or movement acceleration, a context-aware heuristics module can be built to adjust the sigmoid function's probability dynamically. If an AF episode is detected on context change, then subsequent windows can be observed to check whether it lasts more than 50-seconds before making the final decision.

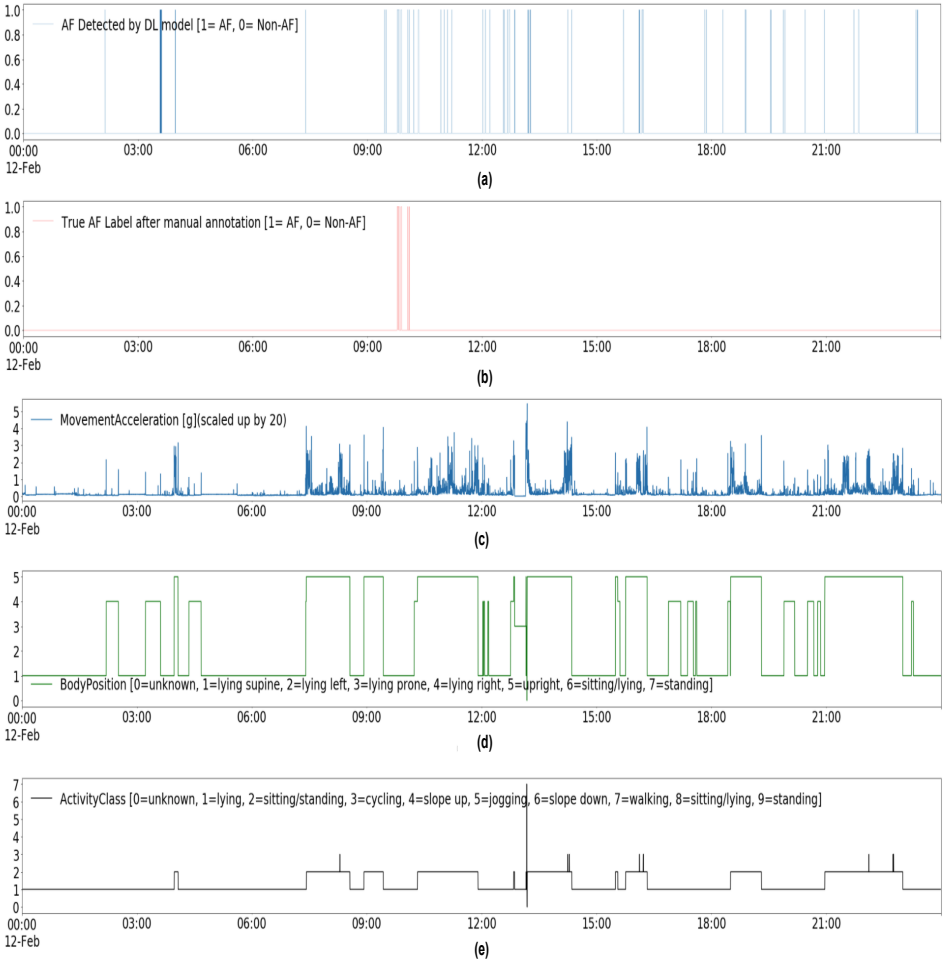


Figure 5.5: User context and false positive occurrences in a 24 hour ECG: (a) shows the AF detected by the DL-algorithm in 24 hours ECG, and (b) shows the ground truth of AF episodes after manual annotation. The short segments ($\leq$ 50-seconds) of FP detected by the DL algorithm in (a) are associated with movement accelerations peaks in Figure (c) and the body position and activity change in Figure (d) and Figure (e) (image from [A.3]).

Table 5.2: Statistics of DL algorithm's performance on CACHET-CADB. AF-DL: number of segments detected as AF by DL algorithm (Table from [A.3]).

Subject	AF-DL	Days	Avg/day	Seg≤50s	Seg≥400s
S1	1991	12	166	1348	68
S2	2826	5	565	1857	25
S3	3419	16	214	1705	317
S4	3712	10	371	1711	107
S5	2308	11	209	1381	166
S6	3198	12	266	1877	65
S7	1702	12	141	1374	32
S8	3058	8	382	2132	153
S9	2415	12	201	1646	55
S10	4290	16	268	2835	71
S11	1236	19	65	883	37
S12	2470	12	205	1707	116
S13	1453	14	103	764	283
S14	787	5	157	614	8
S15	2075	4	518	1152	46
S16	2742	8	342	1569	97
S17	1966	7	280	1435	10

5.4 Summary

This chapter is summarised as follows:

- First, we showed that an algorithm that gives state-of-the-art AF detection performance on public benchmark arrhythmia datasets results in a large number of non-trivial FPs when applied to patient-operated ambulatory ECG under free-living conditions.

- Second, we investigated the impact of ambulatory contexts on FPR in a state-of-the-art AF detection algorithm when applied to ECG from free-living conditions. After testing the algorithm on 215 days of patient-operated ECG from free-living ambulatory conditions, we found that nearly 62% of the total segments marked as AF by the DL algorithm were of $\leq$ 50-seconds, and 99.9% of them were FPs. These 62% of non-trivial FPs segments were mainly associated with three specific user contexts (1) change in activities, (2) change in body positions (especially at night), and (3) sudden movement acceleration. Besides, we found that the TPs were clustered around the early morning and late evening hours. These findings answered RQ 2 and revealed how three specific ambulatory contexts alone induced 62% of non-trivial FPs under free-living conditions.

- Third, based on these findings, we suggest the implication of context-awareness for the design of future DL-based AF detection algorithms. A context-aware heuristics module could be built around these three specific ambulatory contexts. When an AF episode is detected on context change, then subsequent windows should be observed to check whether it lasts for more than 50-seconds. In this way, by dynamically adjusting the sensitivity and specificity around these three FP-prone contexts in free-living conditions, the FPR can be significantly reduced.

CHAPTER 6

CACHET-CADB: Contextualized ECG Database for Arrhythmia Screening

• • •

The content in this chapter is from article [A.4]. It describes the design, development, and validation of CACHET Contextualised Arrhythmia Database (CACHET-CADB). The data collected during the usability and clinical feasibility study of *mCardia* in chapter 4 is used for building CACHET-CADB. The outline of the chapter is as follows. First, section 6.1 describes the motivation behind developing CACHET-CADB. Thereafter, database design, annotation process, and crowd-sourcing tool used for ECG annotation are described in section 6.2. Finally, section 6.3 summarises the statistics of CACHET-CADB and its reuse potential.

6.1 Motivation

In recent years, there has been a significant advancement in the use of machine learning and deep learning (DL) in the field of arrhythmia analysis [40]. The DL models can provide end-to-end arrhythmia detection without requiring manual feature extraction. However, to achieve that, they need to be trained on large ECG datasets. Over the years, many databases such as AFDB [127], MITDB [128], PTB-LX [180], DeepQ [187], CinCDB [31], and OA-ADB [162] have been developed to support this endeavor. The databases such as AFDB and MITDB are some of the oldest and have been extensively used as a gold standard for building and evaluating arrhythmia detection algorithms.

Although the datasets mentioned above have helped move forward the research in computer-aided automatic arrhythmia detection, bringing the models built and evaluated on these datasets into widespread use remains an open challenge for the following reasons:

- The public databases are mostly collected under a clinical controlled environment; therefore, they are relatively clean and have good signal quality [55]. They do not contain all the confounding variations in the signal quality and arrhythmia mimicking artifacts expected in a continuous patient-operated ECG under free-living conditions. The arrhythmia classification models might show high accuracy when trained and tested on the same datasets or other similar clean datasets. However, as shown in the previous chapter, they would result in a large number of non-trivial FP when applied to a patient-operated ECG from free-living conditions.

- ECG signal has significant variations from person to person and is influenced by age, gender, physical conditions, and lifestyle [178, 106, 40]. The existing public datasets are of small size and lack diversity. They are either collected from a single site or a particular type of ECG device. Without having a diverse and multi-site dataset, it is challenging to build a generic arrhythmia detection framework.

- As pointed out in [40, 43], without an understanding of the patient's ambulatory context in which the ECG has been collected, arrhythmia analysis under the free-living conditions remain prone to misclassification. Even in manual analysis of ECG, whenever an ECG segment is inconclusive, physicians usually look for the bigger context or take help from the knowledge of arrhythmia epidemiology [65]. The existing datasets do not provide the patient's ambulatory contextual information during the ECG recording period.

To address the need for contextualized datasets, DeepQ [187] provides the ECG data under three activity types: sitting, walking, and lying down. It was prepared under clinical supervision where participants performed each activity for five minutes while recording the ECG. However, as mentioned earlier, the dataset collected under a controlled environment does not truly reflect the signal quality and confounding artifacts present in the continuous patient-operated ECG under free-living conditions.

Moreover, with the rapid improvement in mobile and wearable technology, wearable ECG devices are becoming more accessible and widespread for longitudinal self-monitoring in patient's natural settings. Thus to train the arrhythmia detection models which can work under free-living conditions, they need to be built and evaluated on ECG data from similar conditions.

To complement existing public arrhythmia databases and address some of the above-mentioned challenges, in this article, we present the design and development of the CACHET Contextualised Arrhythmia Database (CACHET-CADB), a single

channel wearable Holter-based contextualized ECG database. In contrast to existing public arrhythmia databases, CACHET-CADB provides following three unique features:

1. It contains longitudinal continuous ECG data from arrhythmia patients under their *free-living ambulatory conditions*, therefore is most fit for training and evaluating algorithms aimed at enabling real-time ambulatory arrhythmia monitoring in patient-operated ECG.

2. Along with the ECG, it also provides patient's *contextual information* like movement accelerations, activities, sleep, body positions, subjective event (i.e., unusual symptoms experienced during the recording period), and stress levels. This contextual information can make the end-to-end DL-based arrhythmia detection models more explainable (which otherwise remains a black box) and help reduce the false positive detection in free-living conditions.

3. It is multi-site and diverse (collected in Denmark and India).

6.2 Design, development, and validation

The details of CACHET-CADB's design and development process, including ethical consideration, data collection method, data anonymization, ECG annotation process, technical validation, and tools for public use are described in the article [A.4].

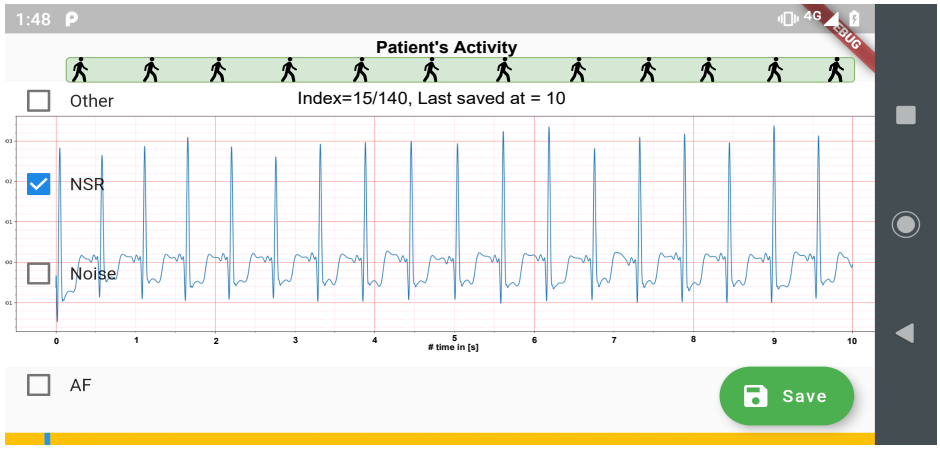


Figure 6.1: ECG annotation tool (image from [A.4]).

6.2.1 ECG annotation tool

Figure 6.2 and 6.1 shows the schematic diagram and UI of the ECG annotation tool used by the cardiologists in manual annotation process. It is built using cross-platform Flutter[1] SDK 1.22.0 and Google Firestore[2] back-end. It presents 10-second sliding window of ECG segments and associated user activities. The annotated data is stored in the cloud Firestore[3]. This tool will be made open-source for crowd-sourcing the manual ECG annotation.

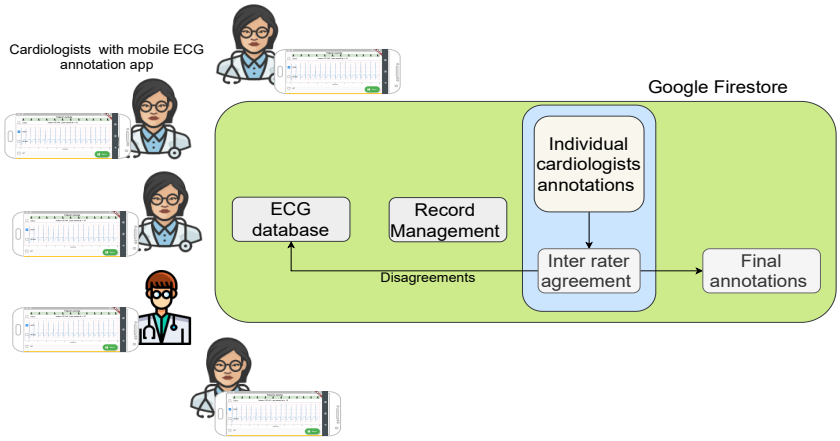


Figure 6.2: Schematic diagram of ECG annotation platform

Table 6.1: ECG annotation statistics in CACHET-CADB. Rhythm class code (1: AF, 2: NSR, 3: Noise, 4: Any other rhythm) (Table from [A.4])

Annotation Class	Number of sample	Rhythm class code
AF	747	1
NSR	615	2
Noise	221	3
Others	19	4

6.3 Contributions and summary

This chapter presented the design and development of the CACHET-CADB, a multi-site, longitudinal, and contextualized ECG database collected under patients' free-

[1] https://flutter.dev/
[2] https://cloud.google.com/firestore
[3] `https://firebase.google.com/docs/firestore`

living conditions. The CACHET-CADB contains 259 days of contextualized ECG data from 24 patients with an average age of 58.83 years. Further, it includes 1602 10-second long manually annotated samples of AF, NSR, Noise, and "Other" (all other) rhythm classes. The statistics of these four rhythm classes are presented in Table 6.1. The CACHET-CADB is made available for download at the DTU Data figshare repository [99] at the Technical University of Denmark. We also designed and developed a mobile-based ECG annotation tool for crowd-sourcing the manual annotation process. The software tools[4] for the reuse and analysis of CACHET-CADB are also made publicly available.

This dataset is particularly important for searchers working on developing reliable and real-time AF detection models, which can work in natural settings irrespective of ambulatory noises and other confounding artifacts. As mHealth based patient-operated monitoring of arrhythmias in a natural setting is gaining traction [162, 28], CACHET-CADB will help in evaluating the models that can work on patient-operated ECG. Additionally, analyzing the output of classification models in context (using CACHET-CADB's contextual data) can make the DL-based end-to-end classification models more explainable and help identify the source of algorithm's error. As no public arrhythmia datasets with contextualized ECG are available, CACHET-CADB could also pave the way for in-depth studies exploring the role of context-aware in personalizing the arrhythmia detection models.

From the perspective of investigating the research question RQ 3 of this dissertation, CACHET-CADB will be used to test if the FP inducing user-contexts found in chapter 5 can be utilized for improving the FPR in an AF detection model. In the next chapter, we introduce *DeepAware*, a hybrid model that combines the DL with a context-aware heuristic model for improving the AF detection under free-living conditions.

[4] `https://github.com/cph-cachet/cachet-ecg-db`

CHAPTER 7

DeepAware: Deep Learning and Context-Aware Heuristic base Hybrid Model for AF Detection

The content in this chapter is from article [A.5]. It addresses the RQ 3 by showing how the awareness of the specific context responsible for inducing the false positives in AF detection algorithm can be utilized to improve the FPR under free-living conditions. The outline of the chapter is as follows. Section 7.1 describes the motivation and design rationale behind building the proposed *DeepAware* model. Thereafter, section 7.2 gives an overview of the architecture and the implementation of the proposed model. Section 7.3 lists the datasets used for building and evaluating the *DeepAware* model and section 7.4 presents the classification results. Finally, section 7.5 summarises the chapter.

7.1 Motivation

Despite good performance on public datasets, bringing DL-based AF detection models in widespread adoption under free-living conditions is challenging due to the high FPR [46]. It was evident in chapter 5 that even though the AF classification model trained on the RR-interval (RRI) features achieved high accuracy (97.6%) and speci-

ficity (89.24%) on AFDB, it resulted in a significant number of non-trivial FPs on single channel ECG from free-living ambulatory conditions. It also resulted in high FPR (13.04%) on MITDB, which has non-AF arrhythmia, and premature ventricular contractions (PVC) beats. They exhibit characteristics of irregular RRIs, similar to that of AF. In addition, the analysis of temporal and contextual distribution of FPs in AF detection algorithm revealed the influence of patient's ambulatory contexts on FPs under free-living conditions.

Based on the recommendations in chapter 5, a new hybrid AF detection model *DeepAware* is proposed to improve the AF detection performance and reduce the FPR on single channel ECG under free-living conditions. To address the non-trivial FP problem caused by patients' contexts (activity change, change in body position, and sudden movement acceleration), *DeepAware* incorporates a new model called context-aware heuristics (CAH) that keeps track of the change in patients' context. Likewise, to reduce the FPR caused by confounding non-AF arrhythmias exhibiting the RRI irregularities similar to AF, the RRI feature-based model (named RR-Net in Figure 7.1) is combined with a DL-based ECG delineation model DENS-ECG [139]. As the P-waves, which represent the atrial depolarization, are absent during AF, the DENS-ECG is used to check the P-waves presence.

By combining RR-Net model with the DENS-ECG model and context-aware heuristics model, *DeepAware* is aimed to reduce the FPR in single channel patient-operated ambulatory ECG. The two main contributions of *DeepAware* over the existing models are its ability to reduce the FPR:

1. In the presence of non-AF arrhythmias exhibiting the characteristics of rhythm irregularity similar to AF.

2. In continuous ECG under the free-living conditions where patient's ambulatory contexts change causes non-trivial FPs.

7.2 Architecture and implementation

Figure 7.1 presents the architecture of the *DeepAware* model. Here a brief description of the architecture components is provided. For details, we direct the reader to the article [A.5]. *DeepAware* comprises of the following six sub-modules: (1) Pre-processing, (2) Segmentation, (3) RR-Net, (4) DENS-ECG, (5) Context-Aware Heuristics (CAH), and (6) Decision box.

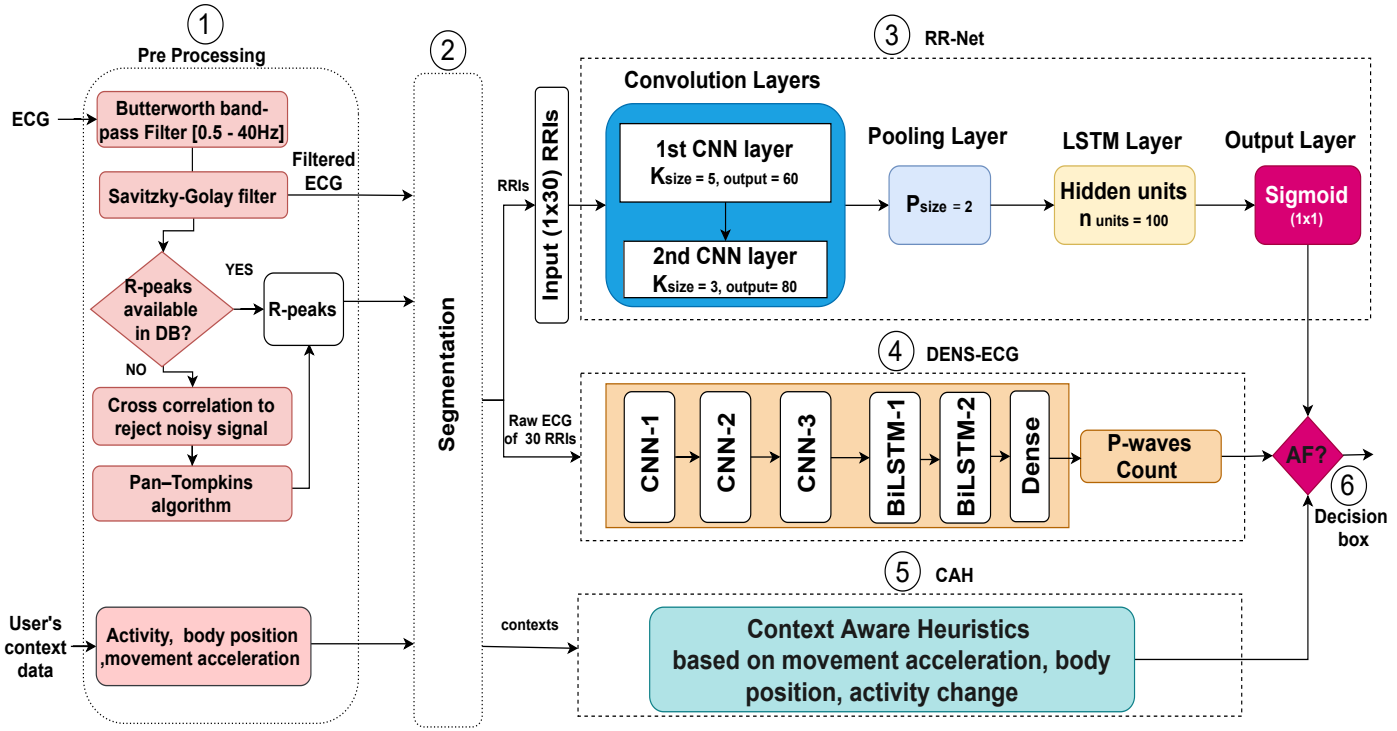


Figure 7.1: The architecture of the proposed *DeepAware* model consists of the following six sub-components: (1) ECG data preprocessing, (2) segmentation, (3) the RR-Net, which take inputs of the RR interval series, (4) the *DENS-ECG*, which takes the raw ECG inputs and gives P-wave count, (5) and a CAH model, which takes user context in a case of ambulatory ECG for dynamically fine-tuning the final output, and (6) Decision box for final binary output (image from [A.5]).

RR-Net: The RR-Net is a combination of a CNN followed by RNN. The CNN is responsible for extracting the features from the RRIs in the convolutional layers. It takes the window of 30 RRIs as input and passes it onto the convolutional layer, which extracts the features. The features extracted by CNN are passed to RNN (a bidirectional LSTM). Thereafter, the sigmoid function finally predicts the irregularity for the i'th RRIs segment, and its probability threshold is set to 0.5. The sigmoid function's probabilities are converted to a binary output of the RR-Net model as follows:

$$\text{RR-Net}(i) = \begin{cases} 1, & \text{if } p(y_i = irregular | \, \mathbf{x}_i, \text{ RR-Net}) \geq 0.5, \\ 0, & \text{otherwise,} \end{cases} \tag{7.1}$$

where RR-Net(i) is the predicted irregularity in the i'th RRIs segment.

DENS-ECG model: DENS-ECG [139] is a deep learning model that combines a CNN and bidirectional LSTM model for ECG delineation to detect onset, peak, and offset of four different components of heartbeat waveforms, namely, the P-waves, QRS complexes, T-waves, and no waves. It is trained separately on QTDB [102]. After the ECG delineation for a window of ECG equivalent to 30 RRI using DENS-ECG, the total number of P-wave are counted, and the binary output for the whole i'th window is defined as per Equation 7.2.

$$\text{DENS-ECG}(i) = \begin{cases} 1, & \text{If } P_{counts} \leq 15 \text{ for 31 heartbeats (30 RRIs),} \\ 0, & \text{otherwise,} \end{cases} \tag{7.2}$$

Context-Aware Heuristics: The context-aware heuristics model is based on the findings of RQ 2 in chapter 5. As it was observed that in an RRI features-based AF detection model (RR-Net here), three specific user contexts, namely change on change in body position, activity change, and sudden movement acceleration, contributed to 62% of the non-trivial FP. The context-aware heuristics module keeps track of context changes in current and previous windows of input RRIs and gives output as per equation 7.3.

$$\text{CAH}(i) = \begin{cases} 0, & \text{if context change detected,} \\ 1, & \text{otherwise,} \end{cases} \tag{7.3}$$

where CAH(i) is the prediction for the context change detection in the i'th RRIs segment window.

Decision box: The output of the above three models RR-Net, DENS-ECG, and CAH is passed to the decision box, which produces the final classification as per equation 7.4.

$$\widehat{D}(i) = \begin{cases} \text{RR-Net}(i) \wedge \text{DENS-ECG}(i), \text{if context not available,} \\ \text{RR-Net}(i) \wedge \text{DENS-ECG}(i) \wedge \text{CAH}(i), \text{otherwise,} \end{cases} \tag{7.4}$$

Where $\widehat{D}(i)$ is the output for the ith window of input.

7.3 Datasets

The technical specification of databases used for building and evaluating the *DeepAware* model are presented in Table I of article [A.5]. The generality of the *DeepAware* has been tested on four datasets – two public and two private. The private datasets are CACHET-CADB (developed in the previous chapter) and CACHET Contextualised Normal Sinus Rhythm Database (CACHET-NSRDB). The CACHET-NSRDB in particular consist of 10 records, each of 24-hours from healthy individuals in NSR. Please note that in *DeepAware* model, the context-aware heuristics module is only enabled/evaluated for CACHET-CADB and CACHET-NSRDB as the other three public datasets do not have patients' contextual data during the ECG recordings.

7.4 Results

Table 7.1 shows classification results of the proposed *DeepAware* algorithms on public datasets (AFDB, MITDB, and NSRDB) and its comparison with other state-of-the-art models from literature. In 10-fold cross-validation on training dataset AFDB, DeepAware achieved specificity, sensitivity, and accuracy of 98.27%, 98.84%, and 98.62%, respectively. The generality of *DeepAware* was tested on MITDB and NSRDB. Compared to [10], *DeepAware* improves accuracy by 4.42% and reduces FPR by 5.63% at the cost of a slight reduction in the sensitivity. Similarly, Table 7.2 shows that compared to just RR-Net (which is trained on RRI features), *DeepAware* improved the accuracy by 3.6% and reduced the FPR by 4.57%. This indicates the importance of incorporating P-wave count via the DENS-ECG model. It helps distinguish AF from other confounding non-AF arrhythmias and ectopic beats, which are falsely detected as AF in models relying only on RRI features. Performance on NSRDB, which has no significant arrhythmias, reveals the expected FPR in healthy individuals with low AF burden. Compared to [10] which is the state-of-the-art model on the NSRDB, *DeepAware* has achieved 3.46% lower FPR.

Performance on CACHET-CADB and CACHET-NSRDB under free living conditions: Table 7.3 shows the *DeepAware*'s performance on CACHET-CADB and Figure 7.4 shows the Confusion Matrix on CACHET-CADB. Compared to RR-Net, *DeepAware* reduces the FPR by 8.07% and improves the specificity at the cost of a slight reduction (1.69%) in the sensitivity. Similarly, *DeepAware*'s performance on CACHET-NSRDB in Table 7.4, which only has NSR recordings from healthy subjects, gives a fair assessment of the expected FPR in healthy subjects under free-living ambulatory patient-operated ECG. The comparisons between the results of RR-Net and *DeepAware* in Table 7.4 demonstrate that the Context-Aware Heuristics model in *DeepAware* was able to significantly reduce the FPR that were induced by the change in patient's ambulatory contexts.

Table 7.1: Performance of the *DeepAware* model on public datasets and its comparison with other state-of-the-art models. Ch: Number of ECG channel (Table from [A.5]).

Algo	Model	Features	Ch	AFDB				MITDB				NSRDB	
				Se (%)	Sp (%)	Acc (%)	FPR (%)	Se (%)	Sp (%)	Acc (%)	FPR (%)	Sp (%)	FPR (%)
[35]	CNN, BLSTM	RRI, Heartbeat Sequences	1	99.93	97.03	96.59	-	-	-	-	-	-	-
[192]	CNN	MFSWT	1	74.96	86.41	81.07	-	-	-	-	-	-	-
[189]	CNN	SWT, STFT	1	98.79	97.87	98.63	-	-	-	-	-	-	-
[105]	CNN	RRI, F-wave frequency spectrum	1	97.4	96.2	97.3	-	-	-	-	-	-	-
[183]	CNN, RCN	Raw ECG	2	94.28	94.91	94.59	-	-	-	-	-	-	-
[193]	MCNN	IHR	2	98.22	98.11	98.18	-	-	-	-	-	-	-
[10]	CNN, BLSTM	RRI	2	98.17	96.29	97.1	3.71	98.96	86.04	87.4	13.96	95.01	4.99
DeepAware	CNN, BLSTM	RRI, Raw EEG, Context	1	98.27	98.84	98.62	1.16	93.05	91.67	91.82	8.33	98.47	1.53

MFSWT–Modified Frequency Slice Wavelet Transform.
MCNN–Multi-Scale CNN.
IHR–Instant Heart Rate Sequence.
RCN–Recurrence Complex Network.

Table 7.2: Performance comparison between RR-Net and *DeepAware* on MITDB and NSRDB datasets (Table from [A.5]).

	MITDB		NSRDB	
Measure	RR-Net	DeepAware	RR-Net	DeepAware
Se [%]	97.74	93.06	-	-
Sp [%]	87.10	91.67	95.53	98.47
Acc [%]	88.22	91.82	-	-
FPR [%]	12.90	8.33	4.47	1.53

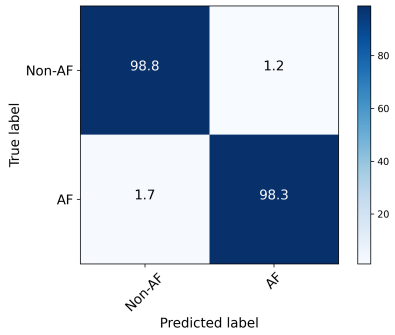


Figure 7.2: Confusion Matrix on AFDB (image from [A.5]).

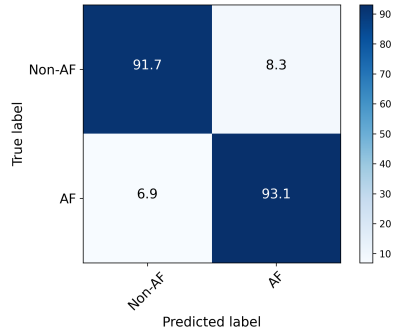


Figure 7.3: Confusion Matrix on MITDB (image from [A.5]).

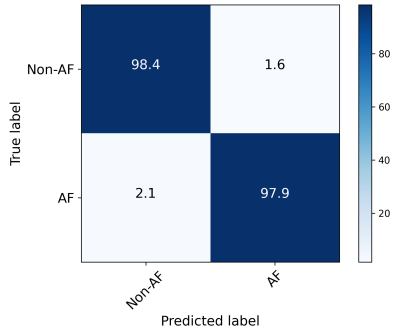


Figure 7.4: Confusion Matrix on CACHET-CADB (image from [A.5]).

Table 7.3: Performance of *DeepAware* on CACHET-CADB (Table from [A.5]).

	CACHET-CADB		
Measure	RR-Net		DeepAware
Se[%]	99.63		97.94
Sp[%]	90.32		98.39
Acc[%]	97.22		98.06
FPR[%]	9.68		1.61

Table 7.4: Performance of *DeepAware* on CACHET-NSRDB. Each record comprising of 24 hours long contextualised ECG under free living ambulatory conditions. Inputs No: Number of (30x1) input windows. (Table from [A.5]).

			RR-Net		DeepAware	
Record	Input No.	r-peaks	Sp	FPR	Sp	FPR
1	5714	114319	89.1	10.9	98.41	1.59
2	5906	118156	88.52	11.47	95.54	4.45
3	3998	80037	89.37	10.63	99.39	0.61
4	3535	70733	91.85	8.15	99.8	0.2
5	1429	28634	97.06	2.93	99.02	0.8
6	4123	82565	82.77	17.23	98.16	1.84
7	5388	108046	95.36	4.64	96.82	3.18
8	5959	119276	80.86	19.13	96.29	3.71
9	4600	92173	94.043	5.95	99.54	0.46
10	5017	100396	95.25	4.7	99.36	0.76

7.5 Summary

In this chapter, *DeepAware* – a hybrid model using the DL and context-aware-heuristics, has been developed and evaluated for AF detection. *DeepAware* beats the state-of-the-art on public datasets. Its generality has been tested on 4 datasets and has shown encouraging results. The use of the P-wave delineation model (DENS-ECG) helped improve the FPR in the presence of confounding arrhythmias that have irregular RRI characteristics similar to AF. Also, on the single-channel ECG from free-living conditions, the effective use of ambulatory context information via context-aware heuristics model has significantly improved the FPR as compared to RR-Net model. Its performance on CACHET-CADB and CACHET-NSRDB show that *DeepAware* can perform real-time AF detection on patient-operated ECG under free-living conditions with relatively lower FPR while keeping the high sensitivity. Through *DeepAware*, we demonstrated the effective use of context-awareness for lowering the FPR and improved the classification accuracy in patients-operated single-channel ECG from free-living conditions. These findings addressed **RQ 3** and partially the main **RQ**.

CHAPTER 8

Discussion

This dissertation was motivated by the potential of using mobile and wearable ECG technology coupled with automated arrhythmia detection algorithms for early diagnosis of arrhythmia under free-living ambulatory settings [140]. From the literature we know that early detection of arrhythmias, and AF in particular, can significantly reduce the massive cost burden on health care and prevent mortalities [82, 140]. Despite the increased capacities of mobile and wearable ECG devices to collect longitudinal ECG and improvements in automatic arrhythmia classification algorithms through techniques such as machine learning and deep learning, their widespread adoption for real-time arrhythmia detection remains an open challenge [140, 40]. Due to the low signal quality of ambulatory ECG, many state-of-the-art automatic algorithms that show excellent performance in clinically recorded datasets result in high FPR when applied to patient-operated single channel ECG from ambulatory free-living conditions [64, 46]. The poor specificity and high FPR in these algorithms could cause over-diagnosis and patient anxiety [28, 96, 140].

Until now, automated arrhythmia detection algorithms have mainly relied on ECG morphology [40, 74] for performance improvement. However, in free-living conditions, context plays a vital role, since ECG morphology changes with the patient's ambulatory contexts [178, 40]. In the absence of patients' ambulatory contexts during continuous ECG, it is difficult at times to ascertain if the ECG morphology changes are due to motion artifacts (i.e., artifacts mimicking arrythmias [116, 51, 114, 138]) or because of a problem with the cardiac conduction system. This adds to the misclassification and over-diagnosis, especially in longitudinal screening. Unlike earlier systems, it is now possible to continuously capture patients' ambulatory context information [133] with mobile and wearable-based ECG devices.

With this background, this dissertation explores whether collecting a patient's ambulatory context information can help improve ambulatory arrhythmia monitoring and diagnosis under free-living conditions:

RQ: *Can contextualized ECG data collected under free-living conditions help improve ambulatory arrhythmia monitoring and diagnosis?*

This question has been addressed in two ways. First, the two clinical cases (see section 4.2.4) demonstrated how and where contextualized ECG helped cardiologists better assess ambulatory ECG. In case 1, the patient's contextual information (sleeping or running) helped in distinguishing SVT from normal heart rhythm, although

both have similar characteristics of heart rate > 100 BPM. Furthermore, in case 2, analysis of the patient's reported event of annoying palpitations in their context (in bed and walking) showed that it could even help understand the patient's symptoms better and assist in making medication adjustments for the arrhythmia patients.

Second, with the *DeepAware* model (see chapter 7), we showed how applying context-aware heuristics helped lower the FPR in an end-to-end, DL-based AF detection algorithm under free-living conditions. The investigation to answer RQ 2 (in chapter 5) revealed that upon applying a state-of-the-art DL-based AF detection algorithm on patient-operated ECG under free-living ambulatory conditions, 62% of all the segments detected as AF by the algorithm were of length $\leq$ 50 seconds, and over 99% (Figure 5.4) of them were FPs. These short segments of FPs were associated with three specific ambulatory contexts: change in activity, body position change, and sudden movement acceleration. This information was used in the *DeepAware* model to build a context-aware heuristics model around these three FPs-prone ambulatory contexts. *DeepAware*'s result on CACHET-CADB and CACHET-NSRDB (Table 7.3 and Table 7.4) demonstrated that combining the RR-Net model with the context-aware heuristics model (in Figure 7.1) improved the model's performance and lowered the FPR on ambulatory ECG from free-living conditions.

Together, these findings demonstrate the usefulness of contextual information in improving ambulatory arrhythmia diagnosis in manual analysis and computer-aided arrhythmia detection algorithms. To address this overall research question, we also addressed three sub-questions. In the following sections, we will discuss our findings and implications for future work for each of these sub-questions.

8.1 Relevant contexts and designing tool for longitudinal arrhythmia screening

RQ 1: *What contextual information is relevant to collect during ambulatory ECG monitoring for improving arrhythmia diagnosis, and what is the design of mobile health technology for collecting such data from patients under free-living conditions?*

With respect to the first part of **RQ 1**, the list of relevant contextual information included an events and symptoms diary, physical activities, and body position, movement acceleration, sleep quality, stress levels, and food intake (see Table 1 in article [A.2]). Factors such as sleep quality, stress levels, and food intake are known for triggering AF in patients [66, 62]. Others, such as physical activities and movements represent contexts that generate noise that mimics AF and other arrhythmias. Please note that these context information categories were elicited based on a thorough literature review and interviews with cardiologists. However, during our analysis and

investigation of the impact of context on false positive rate (FPR), we found some context information was more useful than others. We will further elaborate on this below when discussing RQ 2.

To address the second part of RQ 1, the *mCardia* system was designed for collecting contextualized ECG. The two core challenges in longitudinal arrhythmia screening under free-living conditions are (1) recall bias and frequent non-compliance of the patient-reported symptoms diary, and (2) lack of patient engagement leading to poor signal quality. These core challenges in ambulatory ECG monitoring were addressed in the design of *mCardia*. The feasibility study results reported in chapter [4] and article [A.2] show that the *mCardia* system met its objective of collecting rich contextualized ECG of high signal quality (89% data usable for arrhythmia analysis) while keeping the patients actively engaged in the ECG collection process.

More specifically, *mCardia* improved event reporting practice and reduced recall bias. The ability to report the symptom and event diary through the phone rather than traditional paper-based diary was especially appreciated by the study participants who had previously undergone 1–2 days of ambulatory Holter monitoring at home. As explained by P5,

> *"It was much easier to remember that I had tapped the device and had unusual symptoms by looking at the unfilled event log in the mCardia app. In my previous home Holter test, I rarely maintained the event diary, and even when I did the entries, it was with an approximate time."*

As demonstrated in clinical use case 2 (chapter 4), accurate mapping of these patient-reported events helped the cardiologist to better understand the symptoms and events during ECG analysis. If reporting of these events relies on recall memory and approximate timestamps, symptoms might not be reflected in the corresponding ECG signal during the analysis, which causes ambiguity during analysis.

8.2 Impact of ambulatory contexts on FPR

The second question (**RQ 2**) in this thesis was intended to investigate the impact of ambulatory contexts on FPR in a state-of-the-art arrhythmia detection algorithm when applied to ECG data collected under free-living conditions.

> **RQ 2:** *What is the impact of ambulatory contexts on FPR in a state-of-the-art arrhythmia detection algorithm when applied to ECG data collected under free-living conditions?*

After collecting the contextualized ECG dataset through the feasibility study of the *mCardia* system (in chapter 4), the dataset was used to check the contextual and temporal distribution of false positives when it was applied on a state-of-the-art AF detection algorithm (chapter 5 and [A.3]). The results of the investigation

showed the influence of the patient's contexts on the algorithm's FPR. Specifically, we found that three ambulatory contexts – change in body position, activity change, and sudden movement acceleration – caused nearly 62% of non-trivial FP segments of length < 50 seconds. Surprisingly, it was not the contexts but the change in context (e.g., from sitting to walking/running or sudden movement acceleration) that induced most of these non-trivial short segments of FPR. These relationships may partly be explained by the fact that context change results in irregular RRIs (as illustrated in Figure 6 of the article [A.3]) similar to the those found during AF. However, such RRI irregularities on context change are either heart's natural response to change or due to motion artifacts. For instance, in the transition between sitting to running or walking, there will be a window of RRIs, which has natural variability from the heart. Such natural variability of sinus arrhythmias also falsely gets classified as AF if models rely on only RRI features for AF detection. To overcome such misclassification of sinus arrhythmias as AF, inclusion of atrial activity features is necessary, as employed in *DeepAware*. This will be discussed further below.

The true positive segments in three paroxysmal AF patients were clustered around morning and late evening which corroborate the finding of Hansson et al. [66]. Also, the FPR was slightly higher for female participants than male participants. A possible explanation for this might be that the public datasets AFDB used for training the DL model have an uneven distribution of gender. Also, the breast movements on the chest-mounted Holter might be adding more confounding noisy data for females. As the ECG is known to have significant differences with age and sex [106], the inference from this is that a future model should include more female subjects' data. Taking sex-related ECG differences into account in AF detection algorithms was also recommended by Laureanti et al. [106]. Furthermore, rather than adding more data indiscriminately, future models using the AFDB for detecting AF in free-living conditions should include more data specific to these three context changes in the training set.

Finally, among all the different types of collected context information, only change in body position, activity change and sudden movement acceleration showed strong correlations with FPR. We did not find a strong correlation with other context data such as food, stress-level, location, etc. A possible explanation could be the relatively small recording period (two weeks) or the limited number of participants.

8.3 Contextual information for improving automatic arrhythmia detection algorithm

The third question in this thesis was

> **RQ 3:** *How can arrhythmia detection algorithms be improved by using contextual information obtained under free-living conditions?*

To explore the utility of the contextual information in improving the arrhythmia detection algorithms' performance, there are two possible approaches: (1) using the context information as direct features in the deep learning models, or (2) using them for post-processing heuristics. In *DeepAware*, we have used the context information in the post-processing stage, which aligns with previous recommendations in the literature [149, 137, 65]. Three ambulatory contexts – activity change, changes in body position, and movement acceleration – that influenced the FPR were used to build a context-aware heuristic model. As explained above, in continuous ambulatory ECG, these context changes mimicked the RRI irregularities, similar to AF. For a given input window, if an AF is detected by sub-models of *DeepAware* (RR-Net and DENS-ECG), then the context-aware-heuristics model checked if there was a change in context in the current or previous input window. If the detected AF does not last more than 50 seconds, that is, two consecutive windows, then *DeepAware* changes the final output to non-AF, as it is likely to be a non-trivial FP caused by a change in context. *DeepAware*'s performance on CACHET-CADB and CACHET-NSRDB (Table 7.3, and Table 7.4, respectively) clearly demonstrated that using context-aware heuristics significantly improved the accuracy and reduced the FPR when compared to RR-Net alone on continuous ECG from free-living ambulatory conditions.

Within the scope of this thesis, we have not been able to explore the use of context features as direct input to the deep learning model. However, this could be an interesting line of research to explore given the variety of ambulatory context information that is now available in CACHET-CADB.

8.4 CACHET-CADB for explainable deep learning models

The CACHET-CADB developed in this dissertation complements the existing arrhythmia datasets and provides the additional ambulatory context information during the ECG recording that is lacking in the current public benchmark arrhythmia datasets. In particular, the CACHET-CADB is a valuable resource for the researcher working on developing and validating arrhythmia detection models that can realistically work under free-living conditions on patient-operated wearable ECG.

Although the DL-based models provide the feasibility to build end-to-end arrhythmia classification models, they bring with them the problem of lack of transparency [122]. In the classical feature engineering approach, the clinicians manually selected the features; therefore, it was easy to understand and verify the source of the error in an algorithm. In the case of end-to-end models, features are directly learned by the models themselves; therefore, they remain a black box for the clinicians [122]. As illustrated in Figure 7 in article [A.4], the contextual information of CACHET-CADB can make the end-to-end DL-model output more transparent and explainable, and help identify the source of algorithm error in context.

8.5 Challenges in context-aware arrhythmia monitoring

As demonstrated in this thesis, contextual information has the potential to improve arrhythmia detection under free-living conditions; however, such context-awareness arrhythmia screening brings many other challenges. First, in contextualized ECG collection systems, the patient's privacy and security may be a concern due to the possibility of the misuse of sensitive contextual information [133]. Although our proposed *mCardia* system has security and privacy preservation mechanisms in place, by nature, the continuous collection of mobile context data may give the user a perception of being watched all the time [196]. Second, the continuous context collection drains the battery in mobile and wearable ECG devices. And although the battery capacity of mobile and wearable ECG devices has increased significantly with the improvements in technology, it is still a limitation for context-aware applications that require continuous sensing [196, 133].

8.6 Limitations

The work presented in this thesis is limited in a number of ways. Firstly, the sample size in the *mCardia*'s feasibility study was small, which might limit the ability to generalize the findings. Also, the length of the data collection study was limited to two weeks as it was more focused on technology demonstration or proof of the concept. The usability and patient engagement behavior might change during extended use.

Secondly, the investigation of contextual and temporal distribution of false positives was done on a state-of-the-art AF detection model trained on RRI features. The same might not hold true on other algorithms trained on non-RRI features. However, the models trained on RRI are computationally less expensive and remain the widely used feature in AF detection algorithms [118]. Also, due to resource constraints, the first stage of screening for identifying outrightly NSR or noisy segments was not done by cardiologists. Even though all doubtful segments were passed to the second stage and were reviewed by the cardiologists, there might be a possibility of some misclassification in the first stage, particularly in segments in which the noise and AF mimicked each other. Furthermore, when annotating the samples marked as AF by the DL algorithm, the cardiologists only looked at ECG samples from the onset, offset, and a few random samples in-between to decide the label for the whole segment. It is likely (particularly with long AF segments) that some short ECG samples in-between that were misclassified as AF by the DL model would have gone unnoticed. Nevertheless, this approach was very practical for screening the longitudinal ECG data required for the experiment.

Thirdly, the proposed *DeepAware* is computationally heavy and can work only in a cloud computing environment. In general, the availability of computation, memory, and power on mobile and wearable ECG devices is a significant obstacle for deploying

DL architectures [36]. Furthermore, in *DeepAware*, the DENS-ECG model used for P-wave count has limitations in detecting inverted P-waves, primarily because of the lack of such inverted P-wave morphologies in the training dataset QTDB. Therefore, the overall efficiency of *DeepAware* might differ on the ECG segments with inverted P-waves. In the future, the DENS ECG model used needs to be retrained and tested with more data including inverted P-wave morphology.

8.7 Future work

During the feasibility study of *mCardia*, many false events were registered while taking off the device from the chest or putting it on for charging. In the next iteration, we will resolve this issue by ensuring that event registration by tapping on the ECG Holter can only be done if the ECG Holter is attached. Also, to further reduce the noisy or non-usable data, we will implement signal quality-based alerts for patients and caregivers. In addition, we will address the problem of the ECG device's battery dying, or even being close to dying, going unnoticed when *mCardia* is running in the background, by implementing a server-side notification based on the last known battery status and expected battery life. Furthermore, as we learned during the *mCardia* feasibility study, patients thought that a medication tracking facility would potentially make *mCardia* more engaging. In the next iteration, such a feature could be added to *mCardia* in order to enhance patient engagement in the longitudinal home-based arrhythmia screening.

The long-term plan is to deploy the *mCardia* technology as part of a larger-scale clinical trial in the REAFEL project[1]. This would require that the issues with *mCardia* are addressed and that *mCardia* and the *DeepAware* model are integrated on the back-end server setup. Moreover, a web interface for the clinicians to use would also be needed. During such a clinical trial, the CACHET-CADB will be further expanded.

In the current implementation of *DeepAware*, the contextual information is used in the post-processing stage via the context-aware heuristics and not as a direct feature of the deep learning model. In the future, further research should be undertaken to explore the use of context information directly as an input feature to the DL model.

[1] https://www.cachet.dk/research/research_projects/REAFEL

CHAPTER 9

Conclusion

The objective of this PhD research project was to design and evaluate mobile health technology for ambulatory monitoring and diagnosis of heart arrhythmias. More specifically, we focused on investigating the role of context-awareness obtained via mobile and wearable technology to improve arrhythmia detection on patient-operated ECG under free-living conditions. Two clinical case studies (in chapter 4) demonstrated how and where contextual information helped cardiologists make a better assessment of ambulatory ECG in the manual analysis of a single channel ECG. Furthermore, through the *DeepAware* model, we demonstrated the use of contextual information for reducing the FPR in an automated arrhythmia detection algorithm. Taken together, these findings show that contextualized ECG data collected under free-living conditions can improve both manual analysis and computer-aided algorithms for arrhythmia detection.

To carry out this investigation, we first designed the *mCardia* system that allowed the collection of longitudinal contextualized ECG under free-living conditions. The *mCardia* feasibility study demonstrated its capability for user engagement and collection of quality ECG data needed for longitudinal arrhythmia screening in patient-operated environments. It was also perceived to reduce the recall bias in reporting the symptoms and to improve the patient–doctor communication in the longitudinal screening process. The two clinical case studies demonstrated that context data could provide clinicians with a better perspective of patients' reported symptoms during ECG review.

Using the collected contextualized ECG data, we analyzed the temporal and contextual distribution of FPs in a state-of-the-art end-to-end, deep learning-based AF detection model. The analysis revealed that three specific user contexts, (1) change in activities, (2) body position change, and (3) sudden movement acceleration, were responsible for inducing nearly 62% of the non-trivial FP segments of length $\leq$ 50-seconds. Based on this analysis, we proposed the design implications for future DL models and argued that using context-awareness could reduce the FPs in real-time AF detection under free-living conditions. This analysis also demonstrated that understanding the DL-based, end-to-end AF classification model's outcome in the patient's context can help identify the source of models' shortcomings and make them more transparent.

Based on the above findings, *DeepAware*, a hybrid model combining deep learning and context-aware-heuristics was proposed. In terms of generality, it outperformed

the state-of-the-art AF classification models on public datasets. Its performance on the patient-operated single-channel ECG from free-living conditions (in CACHET-CADB and CACHET-NSRDB) showed its robustness in reducing the FPR in the presence of AF-mimicking confounding artifacts and non-AF arrhythmias. The efficacy of the proposed *DeepAware* model in reducing the FPR on patient-operated single-channel ambulatory ECG from free-living conditions, while maintaining high sensitivity, provides us with the opportunity to use this algorithm "in-house" by cardiologists for longitudinal AF screening. *DeepAware* can significantly reduce cardiologists' workload of manual review of FPs in a clinical setting, allowing them to focus more on treatment than diagnostics.

This thesis also contributed CACHET-CADB, a 259-day-long contextualized ECG arrhythmia dataset from patients in free-living ambulatory conditions. We believe that the CACHET-CADB will help the broader DL community to build and evaluate the arrhythmia detection models that can work under free-living conditions. It will also pave the way for making the DL-based, end-to-end arrhythmia detection models more explainable and help identify the source of algorithm errors, which otherwise remain a black box.

APPENDIX A

Journal Papers

A.1 Mobile and Wearable Sensing Frameworks for mHealth Studies and Applications: A Systematic Review

Authors
DEVENDER KUMAR, STEVEN JEURIS, JAKOB E. BARDRAM, and NICOLA DRAGONI

Journal
ACM Transactions on Computing for Healthcare, Vol. 2, No. 1, Article 8. Publication date: December 2020.

Mobile and Wearable Sensing Frameworks for mHealth Studies and Applications: A Systematic Review

DEVENDER KUMAR, STEVEN JEURIS, JAKOB E. BARDRAM, and NICOLA DRAGONI,
Technical University of Denmark

With the widespread use of smartphones and wearable health sensors, a plethora of mobile health (mHealth) applications to track well-being, run human behavioral studies, and clinical trials have emerged in recent years. However, the design, development, and deployment of mHealth applications is challenging in many ways. To address these challenges, several generic mobile sensing frameworks have been researched in the past decade. Such frameworks assist developers and researchers in reducing the complexity, time, and cost required to build and deploy health-sensing applications. The main goal of this article is to provide the reader with an overview of the state-of-the-art of health-focused generic mobile and wearable sensing frameworks. This review gives a detailed analysis of functional and non-functional features of existing frameworks, the health studies they were used in, and the stakeholders they support. Additionally, we also analyze the historical evolution, uptake, and maintenance after the initial release. Based on this analysis, we suggest new features and opportunities for future generic mHealth sensing frameworks.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Software and its engineering** → **Development frameworks and environments**; • **Applied computing** → **Health informatics**;

Additional Key Words and Phrases: mHealth sensing, mobile sensing, wearable sensing, mHealth frameworks, mobile sensing frameworks

ACM Reference format:
Devender Kumar, Steven Jeuris, Jakob E. Bardram, and Nicola Dragoni. 2020. Mobile and Wearable Sensing Frameworks for mHealth Studies and Applications: A Systematic Review. *ACM Trans. Comput. Healthcare* 2, 1, Article 8 (December 2020), 28 pages.
https://doi.org/10.1145/3422158

This work is partly funded by the Copenhagen Center for Health Technology and the Innovation Foundation of Denmark under grants 6153-00009B and 6152-00005B for projects: REAFEL: Reaching the Frail Elderly Patient for Optimizing Diagnosis of Atrial Fibrillation (http://www.cachet.dk/research/research_projects/REAFEL) and BHRP: Biometric Healthcare Research Platform for Research in Psychiatric and Neurological Diseases using Sensor Technologies (http://www.cachet.dk/research/research_projects/bhrp).
Authors' addresses: D. Kumar, S. Jeuris, and J. E. Bardram, Department of Health Technology, Technical University of Denmark, Richard Pedersens Plads, Building 324, Kgs. Lyngby, DK-2800; emails: {deku, sjeu, jakba}@dtu.dk; N. Dragoni, Dept. of Applied Mathematics & Computer Science, Technical University of Denmark, Richard Pedersens Plads, Building 322, Kgs. Lyngby, DK-2800; email: ndra@dtu.dk.

2637-8051/2020/12-ART8
https://doi.org/10.1145/3422158

1 INTRODUCTION

In the past decade, rapid improvements in processing power, network speed, storage, and the addition of integrated sensors on mobile phones, combined with wearable sensors, have paved the way for new opportunities in research fields such as mobile crowdsensing (MCS), ubiquitous computing, and pervasive health. Today most smartphones are equipped with a wide range of sensing capabilities, such as motion and direction (accelerometer and gyroscope), position (GPS), light intensity, atmospheric pressure (barometer), temperature, proximity, and connectivity (Bluetooth and Wi-Fi). In addition, wearable devices such as activity trackers and medical devices can be connected to the smartphone, which then works as the hub for data collection and processing. Due to the widespread availability of smartphones (nearly 2.5B globally [46]), researchers use them for *in situ* data collection for social, environmental, and mobile health (mHealth) studies. Among others, mHealth in particular has shown growing scientific and commercial interest and has been used for next-generation health research, including *in situ* monitoring and just-in-time (JIT) interventions [2, 10]. Over the years, several mHealth based studies have been conducted in areas such as heart arrhythmia detection, cardiac rehabilitation training, stress, depression, and behavior change [14, 36, 67, 87]. However, to conduct these mHealth studies, researchers may need to implement their own study specific sensing, data storage, and data analysis from scratch, as well as handle non-functional technical challenges such as smartphone resource optimization, configuration of sensors, support for longitudinal data storage and processing, and constant upgrading of the low-level sensing integration to different operating systems (e.g., iOS vs. Android) and smartphone hardware models. All of this requires substantial technical skills and many resources, which is costly and time-consuming. To address these challenges and to support the development and deployment of mHealth studies, researchers have been designing and building more generic and reusable mHealth sensing frameworks. These frameworks provide configurable libraries, modules, and plugins for mobile and wearable sensing, data storage, management, and analysis. Over the years, generic sensing frameworks such as Funf [3], AWARE [24], mCerebrum [35], Beiwe [85], and RADAR-base [68] have been designed and implemented. Such frameworks have typically been released as open source for other researchers to download, adapt, and (re)use. The overall purpose of these frameworks is to help researchers to easily design and deploy mHealth studies with no—or very limited—need for actually programming the sensing technology.

This article presents a systematic literature review of existing mobile and wearable sensing frameworks for mHealth. Such a review is useful for clinical researchers who want to engage in mHealth research studies and are looking for a sensing framework to support their research, as well as for technical researchers interested in existing frameworks to use, extend, or adapt in the design of their own technical solutions and applications. Prior reviews of sensing frameworks have focused on more technical issues such as privacy, resource management strategies, and energy efficiency. For instance, Wang et al. [86] presented a review of various state-of-the-art energy saving techniques in mobile crowdsensing (MCS) applications. Similarly, Christin et al. [16] did a survey on privacy in participatory mobile sensing applications. The review by Khan et al. [42] provided an overview of the state-of-the-art of mobile phone sensing applications in several domains such as traffic and environmental monitoring. Hence, to the best of our knowledge, this systematic review is the first to focus on mHealth sensing frameworks. The review focuses on *generic sensing frameworks* for mHealth, which is distinct from use-case-specific sensing *applications*. Generic sensing frameworks can typically be configured to support several different types of studies (i.e., use-cases/diseases) or can be used to build new sensing applications, whereas use-case-specific sensing applications are tied to a particular study.

This review is conducted as a systematic literature review (SLR) describing the state-of-the-art of currently available generic frameworks that can be used for creating and implementing mHealth studies and applications. The review focuses on describing the functional and non-functional features, historic evolution, maintenance, uptake, and supported stakeholders of the different frameworks. Based on this overview, the article highlights and discusses open issues and potential new features that could be relevant to implement in the further development of such frameworks. The contribution of this systematic literature review (SLR) can be summarized as follows:

- It *identifies*, summarizes, and analyzes existing generic mHealth sensing frameworks and presents the health studies, application areas, and stakeholders each of them support.
- It provides an overview of common *functional* and *non-functional* features in the identified frameworks.
- It examines the *historical* evolution and the uptake of these mHealth sensing frameworks, including how well they are maintained and updated after their initial release.
- Based on this analysis, the review identifies and discusses *open issues* in contemporary research on mHealth sensing frameworks.

The rest of the article is organized as follows: Section 2 describes the systematic review methods, including the research questions, search and screening strategy, inclusion and exclusion criteria, and a description of the data extraction procedure. Section 3 through Section 6 present a detailed analysis of all the functional and non-functional features, health areas and applications, and historical evolution of generic mHealth sensing frameworks. Section 7 provides a discussion of the results, including trends in the research of generic sensing frameworks, their implications for healthcare, as well as new potential features needed in the further development of such sensing frameworks, and potential threats to the validity of this review. Section 8 concludes the article.

2 SYSTEMATIC REVIEW PROCEDURE

In this SLR, we applied the review strategy for conducting SLR in software engineering as proposed by Kitchenham and Charters [43] with a slight modification in search strategy to make it a better fit for this review in the field of mHealth. Specifically, in addition to the proposed search strategy, the backward snowballing technique [90] was applied for finding additional relevant papers. This implies iteratively screening the list of references of the already-included papers in the review.

2.1 Identifying the Need for the Review

Within the past decade, there has been a growing interest in using mobile and wearable technology for collecting contextual, behavioral, and health-related data "in-the-wild," i.e., data that are not acquired in a clinical and/or laboratory setting, but are collected continuously from users in their everyday life. For this purpose, a range of generic mobile-wearable data sampling platforms and frameworks have been designed and released for general use. To search for a systematic overview of existing framework for such mobile and wearable sensing for mHealth purposes, we conducted searches for "systematic literature reviews" in online databases such as ACM, IEEE, Google Scholar, and Scopus using terms such as "mobile/smartphone sensing framework," "crowdsensing frameworks," "sensing framework," and "mobile data collection" in combination with "health" or "behavior sensing" to find any existing SLRs summarizing different generic mobile sensing frameworks for mHealth applications. The search indicated that no overview of this research exists. Therefore, an SLR to summarize and discuss the current state-of-the-art in health-focused generic mobile sensing frameworks is relevant and needed.

2.2 Research Questions

The objective of this review is to identify and provide an overview of all relevant mobile/wearable sensing platforms and frameworks for mHealth and provide an overview of their features (functional and non-functional), which health domain they are used in, and how they have evolved historically and are maintained over time. This overall objective leads to the following four research questions (RQs):

RQ1: Which health-focused generic mobile and wearable sensing platforms and frameworks exist, and which health studies, application areas, and stakeholder do they target?
RQ2: What functional features are supported by these frameworks?
RQ3: What are the non-functional features of these frameworks in terms of extensibility, scalability, security, privacy, license model, and documentation?

Table 1. Number of Papers per Database

Database	Search results
IEEE	330
ACM	226
Scopus	1,058

RQ4: How have these frameworks evolved over time, and how are they reused and maintained after their initial release?

2.3 Search Strategy

The first step of the search strategy was to identify a search string. The search string was constructed by following the guidelines from Khakurel et al. [41] and involved four components. First, we selected keywords used in our previously defined RQs. In addition, we used initial Google Scholar search results to check which keywords are used in popular framework papers. Second, we identified synonyms, acronyms, and alternative phrasings for these keywords. For example, "mobile sensing" for "smartphone sensing," "mobile crowdsensing" for "smartphone crowd sensing," and "behaviour" for "behavior." Third, we merged all synonyms, acronyms, and alternative phrasings using "OR" operations. Finally, all the terms were combined to construct the final search string:

("mobile sensing" OR "smartphone sensing" OR "context sensing" OR "wearable sensing" OR "mobile platform" OR "smartphone platform" OR "mobile context" OR "smartphone context" OR "mobile data collection" OR "sensing framework" OR "data collection platform" OR "smartphone data collection" OR "mobile crowdsensing" OR "smartphone crowdsensing" OR "mobile crowd sensing" OR "smartphone crowd sensing")
AND ("health" OR "human behaviour" OR "human behavior").

In the second step, the first author of this article (DK) used the finalized search string on Dec 20, 2018, to search three electronic databases: (1) IEEE Xplore, (2) ACM Digital Library, and (3) Scopus, limiting the search to publications between 2008–2018 inclusive. The database search was performed on title, abstract, and keywords. These databases were chosen because of their relevance to the field of mobile and wearable technology. Table 1 lists the number of search results for each database.

2.4 Inclusion and Exclusion Criteria

The aim of applying inclusion criterias (ICs) and exclusion criterias (ECs) is to extract only publications relevant to the objective of this SLR. We used the following set of inclusion criterias (ICs) and exclusion criterias (ECs):

IC1: Papers should describe a framework that is generic and supports designing and/or building new sensing applications on top of it, rather than being a use-case or application-specific system.
IC2: Papers should include a reasonably detailed description of the sensing framework's architecture and technical implementation.
IC3: Papers should focus on potential use cases in health, wellness, or behavior sensing.
IC4: Publication date should be on or after the year 2008 (the iPhone was introduced late 2007).
EC1: Papers should not describe a system that is specific to one study or that is not generic enough to support building other mHealth applications than the one described.
EC2: Papers that do not focus on health or well-being.

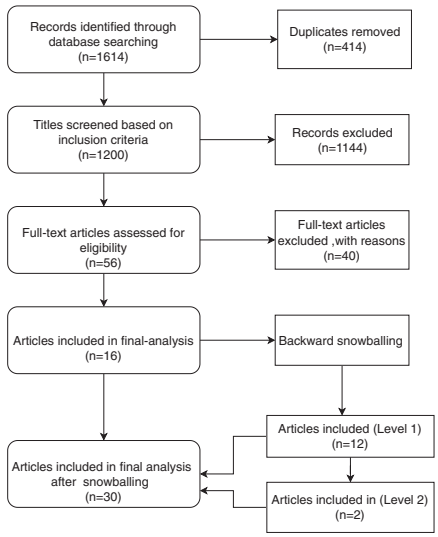


Fig. 1. Review flowchart for selection of papers.

2.5 Screening Process

Figure 1 illustrates the different phases and the number of retained papers in each phase of the screening process. The initial search found 1,614 papers, from which 414 duplicates were removed. The remaining 1,200 papers were then screened based on their titles, adhering to the above-listed ICs and ECs, which removed 1,144 irrelevant papers. Most of these papers were removed due to EC1, since the title often revealed whether a paper presented a use-case-specific sensing application or a generic framework. This resulted in a total of 56 papers for full-text screening, out of which 16 papers were retained for further analysis.

Next, backward snowballing [90] was applied. This implies iteratively looking at the list of references of the included papers and screening them to find new potential papers to include. Through backward snowballing, we identified several papers that used alternative phrasings of "data collection platform" or "mobile sensing frameworks," which we did not anticipate, such as "digital phenotyping" and "social psychology sensing toolkit." Therefore, these papers were only found through backward snowballing, highlighting the effectiveness of this approach. We primarily focused on the references within the "Related Work" sections of the papers. In a first iteration, we included 12 new papers by following references in the initial 16. In a second iteration, we included two new papers. These last two inclusions did not lead to any new identified papers, thereby concluding the snowballing process. Thus, in total, 14 additional papers (12 + 2) were included as part of the snowballing process and at the end of the screening process, 30 papers were included.

2.6 Additional Frameworks

Many of the papers include references to open-source or commercial mobile and wearable sensing frameworks. These frameworks are not published in peer-reviewed scientific venues, but are still similar to the systems found through the systematic literature search and have been widely used for building mHealth applications. To make our review more inclusive, we will also include and discuss these unpublished frameworks. However, given that

they do not have an associated publication, we want to emphasize that their presented data have not been obtained systematically as has been done for the frameworks with a matching publication. Instead, we extracted available information from their website and open-source code base, as available at the time of writing this review. The data tables presented in this review will clearly distinguish between scientifically published frameworks (white background) and unpublished frameworks (gray background).

2.7 Data Extraction, Categorization, and Labeling

The categorization and labeling schemes were created iteratively and were continuously verified through a collaborative process involving all authors. This includes the labeling of application areas, stakeholder classification, sensing typology, and the classification of functional and non-functional features. Full data extraction, labeling, and categorization of the 30 papers was initially done by the first author (DK) and then validated by the second author (SJ). We found two pairs of papers describing the same sensing framework, and one paper from each of these two pairs was removed. Thus, in total, we analyzed 28 distinct sensing frameworks, as listed in Table 2. For information not contained within the papers—such as software license, available documentation, and last update to the code base—we investigated the open-source repositories of the frameworks, if available.

3 FRAMEWORK OVERVIEW, APPLICATIONS AREAS, AND STAKEHOLDERS

This section outlines the results with respect to the first research question (RQ1) introduced in Section 2.2.

3.1 Frameworks

Table 2 provides an overview of all identified frameworks. In total, 28 frameworks have been published in scientific peer-reviewed literature, while 9 frameworks are unpublished (shown on a gray background). In the presentation and discussion of the findings, we will focus on the 28 scientifically published frameworks, unless mentioned otherwise.

We classify 9 frameworks as "end-to-end," as these frameworks provide support for all aspects of running a mHealth study, including data collection and storage, data processing, visualization, participant recruitment, and monitoring study progress. The majority of frameworks ($N = 20$) are built for Android, followed by iOS ($N = 7$) and Nokia ($N = 3$); six frameworks support both Android and iOS operating systems.

3.2 Application Area

Frameworks included in this review are generic and intend to support the design and implementation of mHealth applications in a broad sense. However, to provide an overview of some of the health domain(s) a framework has been used in, we looked at case studies reported in the papers documenting the framework, in papers citing the framework paper, and on the framework's website, if available. Table 3 lists examples of health studies that were implemented using the corresponding frameworks. We find that despite the intention of these frameworks to be general-purpose, most of them still have only been used in a limited number of studies. As shown in Figure 2, the mHealth studies can broadly be categorized into three overlapping categories: *behavioral*, *mental health*, and *physiological health* studies.

In *behavioral studies*, a sensing framework is used to build applications that collect behavioral data (e.g., diet, physical activity). The collected data can be used by behavioral health professionals to understand which behavior contributes to individuals' health condition. Examples include: risky behaviors and HIV transmission using AndWellness [32]; obesity monitoring using HealthOS [48]; StudentLife [87] for assessing behavioral trends, health, and academic performance of college students using Jigsaw [51]; and behavioral correlation between location, social context, and mobility context using AWARE [24]. Out of 28 frameworks, 20 have been used in studies or applications related to human behavioral studies.

Table 2. mHealth Sensing Frameworks and Their Functional Features

Framework				Sensing and Storage						Data Processing, Analytics & Visualization				Study Management	
	Ref.	Operating Systems	Target Stakeholders	Integration with External Sensors	Built-in Smartphone Sensors	EMAs & Surveys	Data Storage & Cloud Back-end	Context-aware Sampling	Remote Configuration	Data Quality Assessment	Data Processing and Analysis	Data Visualization	Behavioral & Health Features	User Consent Support	Study Setup & Monitoring
AndWellness	[32]	A	R, E	✓	✓	✓	✓	✓				✓	✓		✓
AWARE	[24]	A	D, R, E	✓	✓	✓	✓	✓	✓		✓	✓			✓
Beiwe	[85]	A, I	R		✓	✓	✓				✓				✓
CONSORTS-S	[77]		D	✓			✓	✓			✓	✓	✓		
Dandelion	[49]	N	D	✓					✓		✓		✓		
EUPMS	[80]		R	✓	✓	✓	✓								
Emotion Sense	[66]	N	R		✓			✓			✓		✓		
Healthopia	[58]	A	D	✓	✓			✓					✓		
HealthOS	[48]	A	D	✓			✓				✓				
iEpi	[29, 44]	A	D, R	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Jigsaw	[51]	I, N	D		✓			✓		✓	✓		✓		
Lifestreams	[37]	A, I	D, R								✓	✓	✓		
mCerebrum	[35]	A	D, R	✓	✓	✓	✓	✓		✓	✓	✓	✓		
MobiCon	[45]	A	D	✓	✓			✓			✓		✓		
m^k-sense	[31]	A	R		✓	✓	✓		✓	✓		✓			✓
MobiSens	[64]	A	D		✓		✓	✓			✓	✓	✓		
Niima	[5]		R			✓	✓				✓				✓
ODK Sensors	[13, 15]	A	D	✓	✓							✓			
Ohmage	[84]	A, I	D, R, E		✓	✓	✓	✓			✓	✓	✓		✓
Psychlog	[27]	W	R	✓		✓					✓	✓	✓		
QuestionSys	[81]		R			✓									
RADAR-base	[68]	A	D, R	✓	✓	✓	✓				✓	✓			✓
Sensus	[94]	A, I	D, R	✓	✓	✓	✓	✓			✓	✓			✓
SensingKit	[39]	A, I	D		✓		✓				✓				
StarLog	[59]	A	R, E		✓		✓	✓			✓	✓			
TigerAware	[60]	A, I	R, D	✓	✓	✓	✓					✓			✓
UbiqLog	[69]	A	D, E	✓	✓		✓					✓			
Zappa	[74]	A	D	✓			✓				✓				
Bridge	[11]	A, I	D, R		✓	✓	✓		✓		✓	✓		✓	✓
CareKit	[6]	I	D									✓			
Context Sensing SDK	[38]	A, W	D		✓			✓			✓				
Funf	[26]	A	D, R, E		✓	✓	✓	✓	✓			✓			✓
Open mHealth	[57]	A	D				✓			✓	✓	✓			
Passive Data Kit	[82]	A, I	D	✓	✓		✓			✓		✓			
Purple Robot	[72]	A	D		✓		✓	✓	✓						
ResearchKit	[70]	A	D		✓	✓						✓		✓	
ResearchStack	[71]	A, I	D			✓						✓		✓	

The frameworks at the bottom of the table with a gray background are unpublished open-source frameworks. Operating Systems (A = Android, I = iOS, N = Nokia, W = Windows); Target stakeholders (D = Developer, R = Researcher, E = End-user).

Table 3. Health Applications Implemented Using the Frameworks

Framework	Health Studies
AndWellness	Risky behaviors and HIV transmission [32], Behaviors and emotions of young breast cancer survivors [32]
AWARE	Alcohol use events JIT intervention [8], Symptom severity during chemotherapy [50], Detecting drinking episodes [9]
Beiwe	Schizophrenia spectrum illness [85]
CONSORTS-S	Physiological symptom monitoring [78]
Dandelion	Heart-rate and EKG monitoring [49], Fall detection [49]
Emotion Sense	Social psychology [66]
EUPMS	Remote therapeutic interventions [79]
HealthOS	ObeCure: obesity monitoring [48]
iEpi	Study of health behavior in pregnant mothers diagnosed with gestational diabetes [44], Gamified intervention for healthy behavior [44]
Jigsaw	StudentLife: assessing mental health, academic performance, and behavioral trends of college students [87]
Lifestreams	Diet, stress, and exercise in young moms [37], Family wellness study [37]
mCerebrum	Smoking & stress [35], Heart failure [35]
MobiSens	Remote elderly care [91], Mental health monitoring [92]
m^k-sense	Thought and life logging–Mental time travel [31]
ODK Sensors	Heart rate monitoring [15], Diagnose childhood pneumonia [13]
Ohmage	Moms: Studying diet, stress, and exercise-related risk factors for CVD in young mothers [84], PREEMPT: N-of-1 trials using mHealth in chronic pain [84]
Psychlog	Arousal and psychological stress [27]
QuestionSys	Remote therapeutic interventions [79]
RADAR-base	Depression and epilepsy [83]
Sensus	Hourly activity sampling in behavioral activation [73], Social interaction anxiety scale assessment [94]
StarLog	Behavior analysis [59]
TigerAware	Diabetes self-management study [60], Drink and drive [60]
UbiqLog	Mood and sleep [69]
Zappa	Cloud Rehab: Tracking patients with severe brain damage [74]
Bridge	Blood pressure and stress levels tracking [54], Mole Mapper [88]
CareKit	A symptom tracking and reporting instrument mobile application for central nervous system cancer patients [47]
Funf	EMA of day-to-day mood [7], Modeling and discovering human behavior [52]
Open mHealth	Post-traumatic stress disorder (PTSD) [55], Type 1 diabetes self-monitoring case study [56]
Purple Robot	Depressive symptom severity in daily-life behavior [75]
ResearchKit	mPower study (Parkinson's disease) [12], Mole Mapper [88],C3-PRO [65]
ResearchStack	C3-PRO [65]

Mental health studies are built to assess the psychological, emotional, social conditions of human subjects, e.g., to study bipolar disorder, schizophrenia, depression, and anxiety disorder. Although there is overlap between behavioral and mental health disorders, not all mental health disorders are a result of behavioral issues. Of the 28 frameworks, 14 have been used in applications studies related to mental health. Examples include: monitoring schizophrenia spectrum illness using Beiwe [85], and depression and epilepsy using RADAR-base [68].

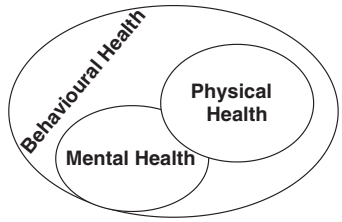


Fig. 2. Behavioral health is a superset of mental and physical health that looks at how behaviors impact health.

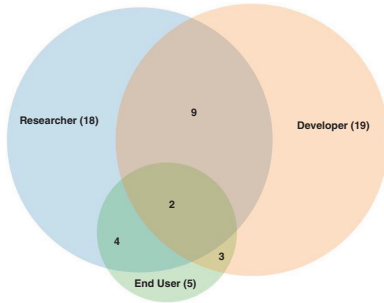


Fig. 3. Number of frameworks supporting different stakeholders.

Physiological health studies involve sensing physiological data such as heart rate, blood pressure, ECG, and skin temperature. Examples include: heart rate monitoring built using Dandelion [49] and heart failure detection using mCerebrum [35]. Of the 28 frameworks, 13 have been used in studies or applications related to physiological health sensing.

3.3 Stakeholders

Since the goal of these general-purpose health sensing frameworks is to support a wide range of health domains and studies with a wide range of potential features, they are also designed to target different stakeholders, depending on their focus. During the labeling of the included papers, we identified the following three categories of stakeholders that the included frameworks provide support for: (i) *researchers*, (ii) *developers*, and (iii) *end-users* including patients. Only a couple of frameworks target all three stakeholders and thereby provide a complete end-to-end solution; most target one or two. Table 2 lists stakeholders per framework and Figure 3 gives an overview of the distribution of frameworks per stakeholder category.

A *researcher*—or study investigator—designs mHealth studies and decides which data to collect to answer a particular research question. They are typically domain experts (e.g., psychologist) and are unlikely to have any experience in software development. This stakeholder group requires support for setting up new studies, setting up personalized interventions, fine-tuning data sampling methods and frequency, triggering surveys remotely, obtaining user consent, recruiting participants, and monitoring progress of ongoing studies. Out of the 28 published frameworks listed in Table 2, two-thirds ($N = 18$) of them target researchers. Examples of such frameworks are QuestionSys [81], EUPMS [80], Beiwe [85], AWARE [24], RADAR-base [68], and Ohmage [84].

Application developers use a framework to implement mHealth applications for data collection and analysis, without having to implement everything from scratch [84]. This stakeholder group expects frameworks to provide secure, modular, and extensible application programming interfaces (APIs) for both mobile phone and server-side development. Two types of frameworks targeting developers can be identified: middleware and back-end frameworks. Middleware frameworks provide application programming interfaces (APIs) to build context-aware applications and pre-processing on phones, whereas back-end frameworks help with storing, analyzing, and visualizing study data on the server side. Nearly two-thirds ($N = 19$) of the 28 frameworks included in this review support developers. Besides, nearly one-third ($N = 9$) of the frameworks support both developers and researchers. All 9 open-source non-academic frameworks target developers and are intended for use in application development (see Table 2).

End-users are individuals or patients who want to use mobile devices to capture health-related data. This is typically done by self-reports (e.g., survey or questionnaires) and continuous sampling from the smartphone's built-in sensors and connected wearable devices. Such users use the sensing frameworks for use cases such as health-awareness and well-being tracking. Since they use their own phones, battery drain is a core concern for them [24, 84]. Such frameworks typically include end-user dashboards to modify which data are collected; e.g., AWARE [24] and Funf [26] support enabling or disabling sensors. Others, such as AndWellness [32], allow end-users to specify a time when to trigger a daily survey. Although most frameworks supporting end-users allow them to view collected data to motivate them for self-reflection and behavioral improvement, only a few (e.g., Reference [32]) support instant feedback. Only Ohmage [84] provides support for privacy control by allowing end-users to delete, export, and change the privacy states of their responses and data. As shown in Figure 3, only 5 out of 28 published frameworks in this review focus on this user group, and only 2 support all three stakeholders.

4 FUNCTIONAL FEATURES

This section describes different functional features supported by the frameworks (RQ2). Based on thematic labeling, the functional features of each framework have been grouped into three overall categories: (i) *sensing and storage*, (ii) *data processing and analysis*, and (iii) *study management*. Table 2 provides an overview of all identified functional features per framework.

4.1 Sensing and Storage

Collection of data and storing it for later access is the primary functionality provided by most of the mHealth frameworks. This overarching feature can be further broken down into support for: (1) *integration with external sensors*, (2) *built-in smartphone sensors and software sensing probes*, (3) ecological momentary assessments (EMAs)*and surveys*, (4) *data storage and cloud back-end*, (5) *context-aware sampling*, and (6) *remote configuration.*

4.1.1 Integration with External Sensors. Wearable sensors are used in many health care studies to enable continuous monitoring of physiology parameters such as heart rate, electrocardiography (ECG), glucose level, sleep, and physical activity. For example, wearables like Fitbit [25] have been widely adopted for monitoring of health symptoms and early intervention in a clinic setting. Nearly two-thirds ($N = 17$) of the 28 frameworks in this review report support for data collection from external sensors. This is typically implemented in one of two ways: through a wired or wireless connection between the smartphone and the sensor, or by retrieving data from the device's data server, where data pre-processing, aggregation, and analysis might be done.

For wireless communication to external sensors, the majority of these frameworks used Bluetooth. Besides, few other wireless protocols such as Adaptive Network Topology (ANT+) and Bluetooth Low Energy (BLE) (in mCerebrum [35]) were also used. Only one framework, ODKSensor [15], provides a plugin interface for connecting external sensors over USB.

Alternatively, rather than retrieving data directly from the sensors, data can be retrieved from the device's remote data repository by calling, e.g., a web API. In this scenario, sensors upload data directly to a vendor's repository either from the sensor or via a vendor-specific smartphone app. Next, vendors may perform additional data-processing on their servers to extract higher-level features that are made available through their APIs. For example, Fitbit performs advanced processing of raw accelerometer data to calculate higher-level features such as step count and sleep patterns. The drawback of this approach is that the framework and its users do not have real-time access to sensor readings, since data are only available through the web API after they have been uploaded to the server. For example, there is a delay of 15 minutes in the case of Fitbit. Additionally, such higher-level features may rely on proprietary algorithms that may change over time. Examples of this type of integration include HealthOS [48], AWARE [24], and RADAR-base [68], which provide plugins to retrieve data from sensor vendors' data repositories. In some frameworks (e.g., AWARE [24]) data collection from the sensor's server is done via the smartphone whereas in other frameworks (e.g., mCerebrum [35] and RADAR-base [68]) this happens server-to-server.

4.1.2 Built-in Smartphone Sensors and Software Sensing Probes. Modern mobile phones have a range of sophisticated built-in sensors that can sense motion, mobility, and the environment, as well as a set of network antennas and communication features, which can be used to sense social and communication behavior. A number of "mobile health sensing" studies have shown that such data can help discover correlations between various physical, behavioral, and mental health conditions [24, 66]. To collect the data from the smartphone's built-in sensors and communication components, mobile sensing frameworks typically provide abstract general-purpose interfaces—often called "probes" [26]. Based on the thematic analysis, support for collecting data from built-in sensors and other components can be divided into five overall categories: (i) motion sensors, (ii) environmental sensors, (iii) communication probes, (iv) network probes, and (v) device probes. Using this categorization, Table 4 provides an overview of the different types of data that can be sensed using built-in sensors and probes on a mobile phone for each of the frameworks included in this review. Please note that these data are based on the available description as provided in the articles and code documentation (if available). It may be the case that only a few relevant sensors were mentioned. Therefore, we do not conclude that omissions in this table imply a given sensor is not supported by the framework. In some cases, articles just stated that the framework supports various built-in sensors but did not provide details. Therefore, it is likely that they might have ✓ under the "built-in smartphone sensors" category in Table 2 but missing details in Table 4.

The most commonly supported sensors mentioned are: accelerometer, GPS, gyroscope, proximity, gravity, light (ambient light intensity), magnetometer, audio, temperature, telephony (start and end time of calls), cell tower (cell towers connected to), Bluetooth (surrounding Bluetooth-enabled and visible devices), and Wi-Fi (e.g., nearby Wi-Fi access points).

Apart from the hardware sensor probes, frameworks also support software and human-based sensing probes. Such software probes include capturing data from the user's calendar, application use, emails, and the call log to capture social activity. The human-based sensing probes include simple prompting of users to input some data (e.g., label an activity for experience sampling method (ESM)) and gesture input [24]. Although some frameworks support multiple operating systems (OSs), the number of supported sensor probes varies from operating system (OS) to OS. For instance, Sensus supports SMS message and light level probes on Android but not on iOS [94]. This is partly due to different hardware setup on the different phones and OSs but also due to iOS implementing a more stringent security and privacy policy, which restricts access to users' personal data such as the phone and SMS logs.

4.1.3 Ecological Momentary Assessments (EMAs) and Surveys. Through the use of surveys, detailed observations or subjective experiences can be obtained by asking users a list of questions. Input requested from users can be in various formats: free text, radio button selection, check-boxes, Likert scale, yes/no queries, and quick responses (a simple button press). ecological momentary assessments (EMAs) involve the repeated collection

Table 4. The Types of Mobile Sensors Each Framework is Capable of Collecting Sensor Data from

Frameworks	Motion Sensors				Environment						Communication				Network		Device		
	Accelerometer	Location	Gyroscope	Proximity	Barometer	Gravity	Light	Magnetometer	Audio	Temperature	Bluetooth	SMS	Telephony	Application	Cell Tower	Wi-Fi	Battery	Screen	Rotation
AndWellness	✓	✓																	
AWARE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Beiwe	✓	✓	✓	✓					✓		✓	✓	✓			✓	✓	✓	
CONSORTS-S																			
Dandelion																			
Emotion Sense	✓	✓							✓		✓								
EUPMS									✓										
HealthOS																			
Healthopia																			
iEpi	✓	✓	✓	✓				✓			✓					✓	✓		
Jigsaw	✓	✓							✓										
Lifestreams																			
mCerebrum	✓	✓	✓		✓		✓	✓	✓										
m^k-sense	✓	✓	✓	✓			✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	
MobiCon	✓																		
MobiSens	✓	✓	✓				✓	✓	✓	✓	✓			✓	✓	✓	✓		✓
Niima																			
ODK Sensors																			
Ohmage	✓	✓							✓			✓	✓		✓	✓	✓	✓	
Psychlog																			
QuestionSys																			
RADAR-base																			
Sensus	✓	✓		✓			✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	
SensingKit	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓						✓	✓	✓
StarLog	✓	✓	✓					✓			✓	✓	✓	✓	✓	✓	✓	✓	
TigerAware		✓																	
UbiqLog	✓	✓					✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓
Zappa																			
Bridge																			
CareKit																			
Context Sensing SDK	✓	✓							✓		✓	✓				✓	✓		
Funf	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Open mHealth																			
Passive Data Kit	✓	✓					✓				✓	✓	✓	✓		✓	✓	✓	
Purple Robot	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ResearchKit	✓	✓	✓					✓	✓										✓
ResearchStack																			

The frameworks at the bottom of the table with a gray background are unpublished open-source frameworks. These data are extracted from the framework's published articles, as well as from their current public code repository/website documentation (if available). The omissions of check-marks (✓) only imply that framework article did not explicitly state if they support that particular sensor probe.

of subjects' current experiences through the use of surveys in real time, in their natural setting. By prompting the user in the relevant context at the right time, EMAs help in reducing recall bias and maximizing ecological validity of the collected subjective data. Half ($N = 14$) of the 28 frameworks report support for EMAs or surveys.

Commonly reported EMA scheduling approaches are: (i) randomized, (ii) triggered at a fixed point in time, (iii) triggered by contextual events, and (iv) remotely triggered by a researcher. The selection of an appropriate EMA scheduling approach is highly use-case specific. Remote scheduling is preferred when a researcher needs to change the survey content very often. When context is important, an EMA can be triggered based on a sensed event, e.g, when the user completes 10,000 steps in a day or when the user approaches a specific geo-location.

mCerebrum [35] implements a bipartite-graph-based EMA scheduler supporting the dynamic adaptation of EMA triggers. It learns from the user's previous responses and adjusts subsequent scheduling accordingly. Sensus [94], AWARE [24], and iEpi [29] support remote and context-triggered EMA scheduling. AWARE [24] employs an EMA questionnaire-building schema defined in JSON, whereas iEpi [29] uses XML to define the content and structure of surveys. Most frameworks do not report on support for multiple survey languages. Only Ohmage [84] supports a few survey languages but not full internationalization.

4.1.4 Data Storage and Cloud Back-end. Frameworks can store the collected data locally and/or on the cloud. In the case of cloud storage, different data offloading techniques are used: when and how often data are synced to remote servers. We found that nearly half ($N = 18$) of the 28 reviewed frameworks rely on cloud storage. Across the frameworks the most common data offloading techniques were: (i) at a fixed point in time, (ii) event-based, and (iii) triggered remotely on demand.

When a fixed point in time for data offloading is chosen, data synchronization with the server takes place at regular intervals. In case the device or server is unreachable, data to be synchronized are queued locally for delivery at a later time. Event-based scheduling is a popular data offloading technique in data collection frameworks that only support low-frequency sensors or surveys. Common examples of events to schedule data offloading are: (1) when a data buffer size reaches a limit, (2) the network connectivity changes from mobile data to WiFi, or (3) the phone is put on charge. Triggering data offloading on demand is typically used in frameworks that involve data sampling across devices. For example, in AWARE [24], the message queue telemetry transport (MQTT) protocol is used to allow devices to issue commands to other devices for data upload, exchange, and synchronization.

SQLite, an embeddable relational database management system, is used as the de facto local data storage for both Android and iOS. However, frameworks that support high-rate sensor data generate large amounts of data. Therefore, such frameworks employ several custom techniques. For instance, mCerebrum [35] implements a custom data router for offloading and efficient data sharing across different applications, called "DataKit." Similarly, RADAR-base [68] relies on Apache Kafka streams for real-time data processing and cloud storage of high volumes of incoming data.

4.1.5 Context-aware Sampling. Context-aware sampling in a sensing framework is when data collection is customized or adapted based on the user's current activity. Nearly half ($N = 12$) of the 28 frameworks report context-aware sampling support to different degrees. Different reasons for using context-aware sampling are: (i) optimizing phone resource (i.e., CPU, memory, battery) dynamics and achieving adaptability based on demands of the application and (ii) to achieve personalization in data sampling by incorporating user's preferences.

Healthopia [58] utilizes context-aware sampling to turn off a subset of sensors based on user context to reduce bandwidth usage and battery consumption (e.g., if the user is not moving then stop measuring the user's heart rate and location). mCerebrum [35], iEpi [29], and AWARE [24] use context inferred from sensors to schedule EMAs. Among the unpublished frameworks, Funf [26], Purple Robot [72], and the Context Sensing SDK [38] provide support for context-aware sampling. In particular, the Context Sensing SDK includes a context API that supports several built-in user context types such as physical activity recognition, location, environment, and

audio classification. It also has a rule-based engine that can be used to create new rules and corresponding event triggers for optimizing phone resource.

4.1.6 Remote Configuration. Remote configuration enables researchers to remotely manage and modify a study at runtime, i.e., after it has been started. For instance, researchers can change the configuration of a running study if they observe that location needs to be sampled more frequently or the timing of an EMA needs to be changed. Only 3 of the 28 frameworks listed in Table 2 describe support for remote configuration. For the others, no description is provided in the paper. Remote configuration is most commonly supported for: (i) changing survey triggers, (ii) adapting survey content, and (iii) changing the sensors' sampling configuration. Frameworks such as AWARE [24] and m^k-sense [31] provide researchers with a web-based dashboard, which enables remote configuration. Similarly, iEpi [29] allows changing sensing and survey parameters in a running study, but does not allow changing contextual triggers for surveys. Amongst unpublished frameworks Purple Robot [72] and Funf [26] support remote configuration. Funf [26] only supports remotely changing the sensors data configuration whereas Purple Robot [72] provides a scripting engine allowing for remote configuration of both surveys and sensor configurations.

4.2 Data Processing and Analysis

Data processing can happen at several levels starting with simple support for *data quality assessment*, to *support for data analysis*, to extracting more high-level *bio-markers, behavioral & health features* from the data.

4.2.1 Data Quality Assessment. Data quality assessment refers to the ability of a framework to assess whether or not incoming data are of acceptable quality for both passively collected sensor data and survey data. This enables warning when data are missing, sensors malfunction or need to be re-calibrated. We found that only a few ($N = 4$) of the 28 articles in our review reported on data quality assessment mechanisms. For passively collected sensor data, they were either used (i) to dynamically adjust depth and complexity of the mobile sensing process or (ii) to alert the researcher about data discrepancies, such as missing data or data outside of expected ranges. Only AndWellness [32] discusses quality assessment of survey data.

As an example, mCerebrum [35] verifies whether incoming high-frequency sensor data are of poor quality and adjusts the depth and complexity by reusing the results from other modules or by choosing an optimal classifier at runtime to achieve optimal performance. An example of supporting data discrepancy alerts is available in the iEpi framework [29], which includes a "compliance report generator" tool, including an analysis of how long each phone was actively collecting data.

4.2.2 Data Analysis. Data analysis tools enable researchers to draw insights from collected data. Depending on the use case, data processing and analysis can be performed either on the mobile phone or on the web server and during (online) or after (offline) the study completed. Nearly two-thirds ($N = 19$) of the 28 frameworks report on their support for data analysis. The granularity and approach for data analysis and processing differs from framework to framework.

The majority of articles report support for post-data-collection data processing (offline) on their back-end servers by providing custom plugins and libraries for data analysis. For instance, Lifestreams [37] is a modular mHealth data analysis stack for personal data sense-making and provides complex feature extraction such as the user's semantic location and physical activity. Similarly, SensingKit [39] describes a plug-in system that allows developers to write automated data processing scripts for pre-processing the data before extraction. Sensus [94] provides an analytics library written in R (SensusR). AWARE [24] does not have ready-made plugins for data processing and analysis, but supports building custom plugins for data processing and analysis. Among the unpublished frameworks, Bridge [11] implements "Synapse," a data analysis subsystem that periodically imports data from the main server for analysis. It creates a separate data analysis environment per individual study to prevent researchers from accessing potentially sensitive data.

4.2.3 Bio-markers, Behavioral & Health Features. In recent years, machine learning and deep learning models have been applied to sensor data to perform real-time behavior feature extraction, activity recognition, and develop bio-markers. mHealth researchers are utilizing such models in applications that require real-time anomaly detection (e.g., continuous stress assessment [36]) and health interventions (e.g., smoking detection [76]). These deep learning–based digital bio-marker models can facilitate new value chains when applied to longitudinal mHealth data, thus help in discovering new behavioral patterns and the personalization of healthcare.

Twelve out of the 28 frameworks support models for bio-markers and health and/or behavior feature extraction. Some prominent supported bio-marker or health feature models are abnormal heart-rate detection [35, 45, 49, 58, 77], fall detection [77], and activity recognition [37, 45, 51, 92]. Among others, mCerebrum [35] has an emphasis on supporting reusable bio-markers and provides models for detecting events such as stress, smoking, and eating. Its architecture enables the creation of study specific bio-marker modules that may build on top of existing ones.

4.3 Study Management

Study management involves handling (1) *participant consent* and (2) *study setup and monitoring*.

4.3.1 Participant Consent. When conducting an mHealth study, obtaining participant consent is essential. The researcher must ensure that study participants are informed about which data are collected, who can access them, and for what purpose they will be used. None of the published frameworks listed in Table 2 provide support for obtaining participant consent, and consent is typically implemented in a study-specific application or obtained in a paper-based form. Among the unpublished frameworks, however, three frameworks [11, 70, 71] provide support for participant consent. Bridge [11] supports defining how consent should be obtained and what it should be obtained for as part of configuring a study protocol, including information such as signature, the scope of data sharing (who all can access data) and consent signing and withdrawal dates. If the consent configuration gets updated during the course of a study, the participant is notified. Likewise, both ResearchKit [70] and ResearchStack [71] provide user interface templates for consent documents with predefined sections such as an overview of the study, information on how data are gathered, privacy policies, how the collected data will be used, time commitment, and consent withdrawal information. Custom study-specific consent sections can be added if needed. Overall, this saves development time.

4.3.2 Study Setup and Monitoring. Nearly one-third ($N = 9$) of the 28 frameworks listed in Table 2 provide support for study setup and monitoring through a user interface, typically through a web portal. This includes support for: (i) study participant recruitment, (ii) creating and scheduling surveys, (iii) visualizing study progress, and (iv) remote configuration of sensors and survey triggers.

All the frameworks that support study setup and monitoring provide a dashboard for the researchers to visualize the collected data. Besides this, frameworks such as AWARE [24], Beiwe [85], and Sensus [94] provide assistance in registering study participants, creating, editing, and deploying surveys. AWARE [24] additionally supports editing sensor sampling configurations, changing the database, adding co-researchers, and viewing the devices linked to a study. In addition to providing researchers access to collected data, two frameworks—Ohmage [84] and Andwellness [32]—also enable study participants to view their data. Ohmage [84] in particular further allows the participant to change their privacy settings, delete, and export their study data.

5 NON-FUNCTIONAL FEATURES

This section presents the non-functional features of the frameworks and hence addresses the second part of RQ2. In traditional software engineering, a non-functional requirement (NFR) is defined as "a software requirement that describes not *what* the software will do but *how* the software will do it" [1]. non-functional requirements (NFRs) are used to evaluate the operation of a system, rather than specifying its functional behavior. For the 28

published frameworks, we looked for NFR details in the papers. For the unpublished frameworks, we looked at information on their respective websites, but overall very limited information was available. Please note that in case no details on an NFR are presented, this does not necessarily mean it is not addressed by the framework, but simply that we could not find any mention of it in the screened sources. The review focused on four categories of NFRs, which are most relevant to mobile sensing frameworks: (i) *extensibility*, (ii) *scalability*, (iii) *security and privacy*, and (iv) *license and documentation*. An overview of all identified NFRs per framework is provided in Table 5.

5.1 Extensibility

In software engineering, extensibility denotes the ability to add, enhance, or repair existing functionality in a system or component [19]. Since the number of available built-in sensors in modern smartphones keeps increasing and new wearable sensors emerge continuously, it is important that these can easily be integrated into existing sensing frameworks. For extensibility, we checked if the framework article describes an APIs for adding new functionalities, including support for adding new sensing capabilities. More than two-thirds of the articles ($N = 19$) state that they support extensibility. It is interesting to note that a total of 10 articles just state support for extensibility without providing any further details. In the articles describing the extensibility mechanisms, there were typically two ways to extend the framework: (i) an API for adding new smartphone sensing modalities (such as a light sensor on the phone) and/or data types to be collected or (ii) a plugin mechanisms for adding support for external sensors (like an electrocardiography (ECG) monitor).

Examples of frameworks supporting the first type of extensibility include AndWellness [32], AWARE [24], Passive Data Kit [82], Funf [26], and Purple Robot [72], which all support extensibility and modifiability by providing an interface to add new custom smartphone sensor probes. Extensibility can also be supported on the back-end of a framework, where extensibility focuses on the ability to support new data types of custom data sources. For instance, Ohmage [84] implements a backward-compatible API to achieve extensibility. However, Open mHealth [57] allows adding new custom data schemas and also provides an extension for integration with external data sources such as HealthKit and data collected from electronic health records (EHR)s.

In the second category, we identified frameworks such as AWARE [24], UbiqLog [69], and HealthOS [48], which support extension of the framework to collect data from external sensors. ODK Sensors [15] supports extensibility by providing an interface between external sensors and a smartphone for managing device discovery, data buffers, and communication channels. ODK Sensors also provides an ecosystem of underlying reusable sensor drivers that enables developers to write minimal sensor-specific code.

5.2 Scalability

In software engineering, scalability is a desirable property for a computing system to have. And while the basic notion is intuitive, scalability has no generally accepted definition [33]. Therefore, in this review, we examined how scalability is described in the articles or websites of the included frameworks. Of all the frameworks listed in Table 5, more than two-thirds ($N = 19$) of the articles mention support for scalability. However, out of these, only 7 present additional details to back up the claim that the presented framework is scalable. For each framework that provided details on scalability, we investigated whether support for scalability was (i) argued based on the underlying technology used or (ii) claimed based on a dedicated scalability evaluation.

In two articles, the authors argue for the scalability of the framework based on a description of the underlying infrastructure and technology used to implement the framework. Ranjan et al. [68] state that RADAR-base relies on the Apache Kafka platform [20] as the underlying infrastructure to achieve scalability. However, no scalability evaluation is reported. In AndWellness [32], the authors state that they rely on the Spring framework for server scalability, as it provides component replaceability and flexibility.

Some articles report on scalability evaluations. Examples include mCerebrum [35], AWARE [24], and Ohmage [84]. The article on mCerebrum [35] describes a high-frequency sensor data collection test on the phone, the results of which indicate that the framework outperforms others, such as GoogleFit [28], AWARE [24],

Table 5. Non-functional Features

Framework	Extensible	Scalability	Security		Privacy	Open	Docs
AndWellness	✓	✓	✓	DENC,SPC	✓		
AWARE	✓	✓	✓	DENC,SPC	✓	✓	✓
Beiwe	✓	✓	✓	DENC,EXT,SPC	✓	✓	✓
CONSORTS-S	✓						
Dandelion	✓						
Emotion Sense		✓				✓	✓
EUPMS			✓	DENC			
HealthOS	✓	✓	✓	DENC	✓		
Healthopia					✓		
iEpi		✓	✓	DENC	✓		
Jigsaw	✓	✓					
Lifestreams	✓				✓	✓	✓
mCerebrum	✓	✓	✓	DNA	✓	✓	✓
MobiCon	✓						
MobiSens					✓		
m^k-sense		✓	✓	DENC	✓		
Niima	✓	✓	✓	DENC,SPC	✓		
QuestionSys		✓					
ODK Sensors	✓	✓					
Ohmage	✓	✓	✓	DNA	✓	✓	✓
Psychlog						✓	
RADAR-base	✓	✓	✓	DENC	✓	✓	✓
Sensus		✓	✓	EXT	✓	✓	✓
SensingKit	✓					✓	✓
StarLog					✓		
TigerAware	✓	✓	✓	DENC	✓		
UbiqLog	✓	✓	✓	DENC,SPC	✓	✓	
Zappa	✓	✓				✓	
Bridge	✓	✓	✓	DENC,SPC,EXT	✓	✓	✓
CareKit	✓	✓				✓	✓
Context Sensing SDK	✓				✓		✓
Funf	✓	✓	✓	DENC	✓	✓	✓
Open mHealth	✓	✓	✓	SPC		✓	✓
Passive Data Kit	✓					✓	
Purple Robot	✓					✓	
ResearchKit	✓	✓				✓	✓
ResearchStack	✓	✓	✓	DENC		✓	✓

The frameworks at the bottom of the table with a gray background are unpublished open-source frameworks. Check marks (✓) indicate that the article provides a description of the listed non-functional feature. Open = Open Source, SPC = Secure protocol, EXT = Rely on external infrastructure for security, DENC = Data encryption, DNA = Details not available (the framework states that it has support for security, but details of the security mechanisms are not specified).

and HealthKit [30]. Other articles report on scalability tests on the server-side. For example, the Ohmage [84] server was load-tested using over a million API calls, and scalability of the AWARE [24] server dashboard was demonstrated by increasing the number of participants and measuring response time for page load, visualization, sorting, and searching.

5.3 Security and Privacy

Since sensitive health and behavioral data are collected as part of mHealth studies, support for security and privacy in sensing frameworks is essential. In this review, we have investigated which techniques the frameworks use to ensure *data security* as well as *privacy*.

5.3.1 Security. Surprisingly, we found that only half ($N = 14$) of the 28 articles discuss security techniques. The main security techniques presented are (i) data encryption (DENC) and (ii) data transfer over secure protocols (SPCs).

Data encryption is a security method that allows data encoding that then can be only decoded with the correct encryption key. A total of 11 articles in this review report using encryption for security, and five out of these stated using Advanced encryption standard (AES). Advanced Encryption Standard (AES) is encryption specification set by United States' National Institute of Standards and Technology [63] for encryption of electronic data. AWARE [24], Beiwe [85], and Ohmage [84] are examples of frameworks that use AES. Among the unpublished frameworks, three frameworks use encryption: Funf [26], Bridge [11], and ResearchStack [71].

All frameworks that provide details on how they securely transfer data to the server (SPC) do so by using Secure Sockets Layer (SSL). Among unpublished frameworks, two frameworks—mHealth [57] and Bionetworks [11]—use Secure Sockets Layer (SSL). In addition, there are three frameworks in which the authors state that they provide secure data collection and offloading, but do not provide any additional details (labeled as "DNA" in Table 5).

Three frameworks—Sensus [94], Beiwe [85], and Bridge [11]—report that they rely on the security mechanisms of a cloud provider (e.g., Amazon Web Service (AWS)) for back-end data security (labeled as "EXT" in Table 5).

5.3.2 Privacy. Collecting contextual and physiological data in mHealth sensing involves significant privacy risks. Therefore, it is relevant for mHealth sensing frameworks to implement privacy-preserving techniques as part of data collection, processing, and sharing. We found that 17 of the 28 articles discuss privacy-preserving techniques. The key techniques used are: (i) deleting personally identifiable information, (ii) one-way hashing, and (iii) storing personal information and sensor data on separate servers.

One-way hashing is a technique that converts a message into a unique message digest in an irreversible way. It is often used to obfuscate and protect sensitive information. Three frameworks [5, 24, 85] use one-way hashing to obfuscate personal identifiers (e.g., phone numbers and MAC address). Beiwe [85] uses the standard SHA-256 algorithm for hashing phone numbers, whereas Niima [5] uses a two-stage hashing process in which the first hashing takes place on device and the second hashing is done on the server using a secret Salted Hash [40] algorithm. One framework—RADAR-base [68]—keeps sensor data and user information on separate servers to achieve pseudonymization. mCerebrum [35] introduced a new privacy preservation technique by means of implementing central privacy and access controllers. Its central privacy controller allowed participants to suspend sensor specific data collection and sharing.

Among the unpublished frameworks, both ResearchKit [70] and ResearchStack [71] provide means to list privacy policies in the user consent module. However, the implementation details of the specific privacy technique is left to the individual applications that are built using these frameworks. Two others—Funf [26] and Bridge [11]—use one-way hashing for anonymization.

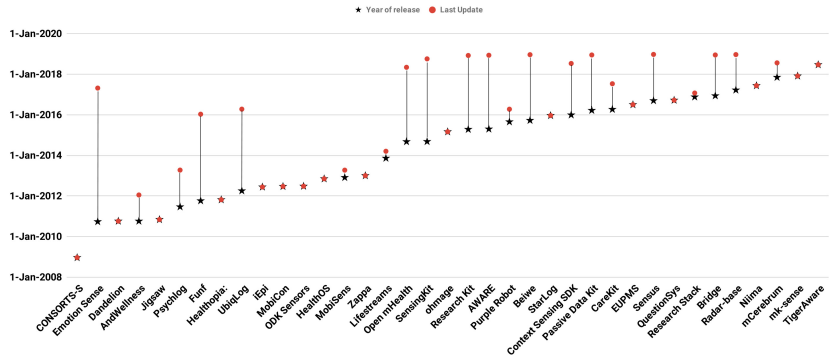


Fig. 4. Initial release and last update time for all frameworks.

5.4 License Model and Documentation

To make sensing frameworks reusable, maintainable, and extendable, detailed and accurate documentation is necessary. Moreover the software license model is important for developers and researchers to consider when choosing a framework, since copyright restrictions could limit how the framework can be used and modified. To obtain information about the license and available documentation, we looked at framework websites and their code repositories in the case of open-source frameworks.

5.4.1 License Model. Over one-third ($N = 12$) of the 28 frameworks in this review are open source. Seven of them are licensed under the Apache License 2.0, three under BSD 2-Clause "simplified," one under BSD licenses, and one under GNU Lesser GPL v3.0 (LGPL). All the unpublished frameworks except for Context-Sensing SDK [38] are open source: five are licensed under the Apache License 2.0., two under BSD License, and one under GPL v3.0.

5.4.2 Documentation. Since it is difficult to assess whether a framework is sufficiently documented, we only looked at whether documentation of any kind, be it on an architectural level or API level, is provided. We found that less than one-third (N = 9) of the 28 frameworks provide some level of documentation. In contrast, all of the unpublished (mainly commercial) frameworks provide detailed API documentation.

6 MAINTENANCE

The main purpose of researching, designing, and implementing a mHealth framework is to support re-usability of core components and technology, thereby enabling easier creation of mHealth applications. This kind of re-usability and cross-application technology requires, however, that the frameworks are updated and maintained on a regular basis to accommodate changes in the underlying OS and to support new features. This section investigates the maintenance of the frameworks, i.e., research question 4 (RQ4). We are interested in understanding how the identified mHealth sensing frameworks have evolved over time and to what degree the frameworks are being updated and maintained after their initial release.

To identify the period over which frameworks have been maintained, we compare the first and last date at which information about the framework has been made available, illustrated in Figure 4. Concretely, for the first date, we used the publication date of the framework's article or the date of the first release of the software for non-published frameworks. For the last date, we chose the date of the last commit to the framework's public code repository or alternatively the date of the last listed release on the corresponding website. We found that more than two-thirds ($N = 21$) of the included frameworks (including the unpublished frameworks) maintained

their repositories after the initial release. Ten frameworks have been updated in the year 2018 (the time this review was concluded). AWARE [24], SensingKit [39], Sensus [94], RADAR-base [68], Beiwe [85], Bridge [11], and mCerebrum [35] are some of the most regularly updated frameworks that were still being maintained in 2018. The past few years (2015–2018) have seen a dramatic increase in the number of released frameworks. Also notable is that up until 2014 most frameworks were middleware libraries, whereas most recent frameworks (in or after 2015) are end-to-end mHealth frameworks that include support for sensing, storage, processing, visualization, and study management.

7 DISCUSSION

This review has shown that, since 2010, dedicated research into mobile and wearable sensing frameworks for mHealth has resulted in a non-trivial set of sensing technologies. Approximately ten frameworks are maintained today (see Figure 4) and the more recent ones are far more comprehensive, including sophisticated functional and non-functional features on both the phone and the server side of the technology (see Tables 2 and 5). In addition to the early frameworks that were mainly driven by academic researchers there has been a growing interest in mHealth from industry. Several commercial frameworks, such as HealthKit [30] and GoogleFit [28], have been introduced to support the creation of mHealth sensing applications in the Apple and Google technology ecosystems, respectively.

In terms of functional features, several recent frameworks have focused on functionalities such as remote configuration and study monitoring support (see Table 2), which enable researchers to change data sampling configuration, survey content, and triggers remotely. Also, given the advancement in machine learning and deep learning, support for features such as behavioral and health-biomarkers computation is starting to be included in recent frameworks [35]. Due to limited resources on the smartphone (e.g., computational power, battery, and network bandwidth), the initial work on feature extraction and data analysis was mostly done offline on servers. Adding support for high-level feature extraction on smartphones has the potential of enabling the design of advanced just-in-time interventions in health sensing applications.

The unpublished open-source frameworks seem more popular, gauged by the number of people that starred their code repositories. This might be due to the fact that these frameworks typically come with good documentation, including tutorials, and are maintained by a professional software development team of a large company. Amongst the research-based frameworks, AWARE [24], Sensus [94], and Beiwe [85] seem to be the most popular in terms of people that starred their repositories.

The majority of framework evaluations focused on system resource management and impact on battery life while sensing, storing, and transferring data over the network. However, the evaluations did not use a common or standardized way to evaluate performance and efficiency. For example, different versions of smartphones from different manufactures were used. This complicates comparing efficiency and performance across frameworks. Nevertheless, there are some recurring findings. Evaluations of frameworks such as StarLog [59], AndWellness [32], SensingKit [39], and Aware [24] show that configuring high sampling frequencies for sensing probes is associated with draining the battery very rapidly. But, there are ways to mitigate this. An evaluation of mCerebrum's micro-batching strategy showed a significant reduction in CPU use while ingesting high-frequency sensor data. Thus, this makes micro-batching a good strategy for applications that require real-time data sharing and processing. In addition, none of the reviewed frameworks reported on health efficacy of applications built and deployed using them. They solely report on technical proof-of-concept studies within different health domains. But, it must be noted that such studies are typically reported in separate papers, referring back to the toolkit/framework papers, which were not included in this review.

7.1 Implications on Healthcare Research

There is mounting evidence that mHealth applications are practical and low-cost means to deliver disease prevention, self-management, diagnostics, and treatment based on commodity hardware. A simple app could give

anyone, anywhere, the ability to perform diagnostic screening tests and monitor their condition after they receive an official diagnosis [53]. Moreover, healthcare may become more available and affordable in resource-poor settings in a way that was previously unimaginable [4, 22]. The reviewed frameworks enable the creation of such mHealth apps in a manner in which developers and researchers can more easily design, implement, deploy, and monitor mHealth applications and studies for a growing number of diseases and user groups. As shown in Table 3 and Figure 3, the reviewed frameworks have been utilized for building a diverse range of mHealth applications for different stakeholders.

However, the review also points to a set of issues in this line of research. First, despite that a major part of the frameworks were designed for mHealth application development and targeted developers as their primary stakeholders, we found that such frameworks were actually only used in a limited (1–3) number of mHealth applications. And, they are rarely used by researchers outside of the group of researchers who implemented the framework in the first place. Hence, it seems that the overall goal of having application developers adopt reusable mHealth frameworks has still not been achieved.

Second, from a healthcare perspective, the use of detailed data sampling from mobile and wearable devices still needs to fit into the overall clinical treatment and care of diseases. Many of the application studies reported are still early work and basic research, trying to understand the role and utility of mHealth sensing. Much of the research so far has targeted sensing from the mobile phone—such as accelerometer, movement, and phone interaction—that reveals information about the patient's behavior but very little about the patient's health status. Very limited work has been addressing general-purpose collection of health-related data such as physiological signals, bio-chemical data, and medication. A few frameworks—like HealthOS [48]—provide custom adapters to medical wearable devices, but their adapters are not generic and are limited to a handful of consumer health devices (i.e., not medical grade). Moreover, due to the voluminous size of data and limited time of researchers, not all insights that could be obtained from the collected data are found [17, 89]. Thus, there is a need for automatic detection of higher-level biomarkers for these frameworks to be useful in the clinic. Hence, from a healthcare perspective, multi-parametric data collection as well as the recognition of higher-level biomarkers will be needed.

Third, the frameworks available now are quite mature and seem to be maintained and backed by larger groups or consortiums of researchers, rather than individuals. This is promising for the use of these frameworks in the design and development of mHealth applications for clinical use. However, the main focus of current frameworks is on data collection, which is useful for population screening and disease diagnosis and monitoring. But, there is little or no support for interventions and delivering treatments. Examples of such treatment components could be targeting educational material, medication management including reminders, behavioral or cognitive therapy, and recommendations for healthy living. As such, the flow of data and information in these frameworks seems to be "one way," i.e., flowing from the patient to the clinicians, while little or no information flows in the other direction. Some of the frameworks are, however, designed to be used in conjunction with other frameworks, which may support interventions. But for the majority of the frameworks this is not the case.

Nevertheless, for healthcare research, the use of such data sampling frameworks may open doors for large-scale data analysis in healthcare. "Big data" and cross-study analysis can help researchers in unearthing new trends in disease management and treatments at both the individual and population level. Also, through standardized and exportable study protocols (sampling profile, data analysis tools, survey, and EMAs configurations, etc.) frameworks could significantly enhance study reproducibility, replicability, and transparency in mHealth research.

The selection of a framework for research or development purposes should be done with care. We hope this review can help guide healthcare researchers and/or mHealth application developers in the selection of a framework that suits their needs.

7.2 Research Gaps and Recommendations for Future Research

This review has shown that several mHealth sensing frameworks exist, some of which are maintained and are fairly popular in terms of use and offer quite an advanced set of functional and non-functional features. As such, research into mHealth sensing frameworks seems to have come a long way, and several frameworks are available to choose from for researchers and application developers. A relevant question is then whether there are still open research questions to be addressed. Based on our review, we have identified six gaps and thus opportunities for further research. This section discusses these gaps and provides recommendations for future mHealth frameworks.

7.2.1 Personalizing Data Collection. Frameworks suffer from poor power management and cloud offloading strategies [35, 92]. In most frameworks, sampling frequencies, data processing, and data offloading techniques are either hard-coded or one configuration is provided for all participants. The existing frameworks do not take individuals' preferences, mobile hardware, resource availability, and so on, into account. Such non-personalized data collection results in poor resource management, which can be expected to annoy users of longitudinal mHealth studies. Even though frameworks such as AWARE [24] and Beiwe [85] support remote re-configuration of sampling probes after the deployment of a study, this reconfiguration needs to be done manually and cannot be personalized to individual participants. We argue that resource management and therefore user experience in mHealth applications can be improved by incorporating adaptive personalization and dynamic re-calibration of sampling profiles. This could include the use of various code offloading techniques—techniques that dynamically migrate processor-intensive tasks to cloud surrogates.

7.2.2 Data Standardization. The frameworks described in this review support data collection from a wide range of sensors. These data, however, are not stored in any standardized format and none of the frameworks that do sensing support specifying standardized data schemes to use for sampled data. This implies, for example, that blood glucose measurements collected from the same device but via two different frameworks cannot be directly compared, as there is no common standard for storing blood glucose (including the context within which the measurement was collected). Similarly, since patient-generated data do not comply to any clinical standards, these data are hard to integrate into existing clinical systems, limiting its use [89]. Furthermore, to allow for reusable higher-order features and bio-markers extraction, and for cross-device and cross-study analysis, it is essential that all the data points for a specific measure (e.g., heart rate or blood glucose) are represented in a standardized format, regardless of the framework and/or devices from which the data were collected.

Currently, there are some initiatives to standardize mHealth data. The Health Level 7 (HL7) standardization initiative has a "Mobile Health" work group, which creates and promotes health information technology standards and frameworks for mobile health [34]. The Institute of Electrical and Electronics Engineers (IEEE) develops and maintains the IEEE 11073 standards on "Personal Health Devices' [61] as well as the P1752 open standard for Mobile Health Data [62], which is based on the Open mHealth initiative [57]. However, none of the reviewed frameworks (except Open mHealth of course) supports any of these standards. Therefore, this could be a topic for the evolution of mHealth frameworks.

7.2.3 Measurable User Privacy Support. Mobile and wearable sensing frameworks provide a novel opportunity for long-term health sensing. However, due to the sensitive nature of health data, such frameworks pose a risk to users' privacy, which is important to consider, especially given the recent introduction of the General Data Protection Regulation (GDPR) in Europe. Even though several frameworks such as mCerebrum [35] and AWARE [24] have implemented various privacy protection mechanisms (e.g., pseudonymization; central privacy and access controllers), they are mainly focused on privacy protection as part of data collection. But, when sharing data across different studies to enable "big data" analysis, additional concerns arise. Data should not be used for a different purpose than the user has given consent for. Similar to security and encryption standards,

incorporating privacy compliance standards (e.g., BS 10012:2017 [23]) in mHealth frameworks could help in achieving measurable user privacy.

7.2.4 Measurable User Feedback. We observed that out of their three target audiences, namely, developer, researcher, and end-user, frameworks in this review mainly reported on the end-user's feedback and usability evaluation. Except for a few (e.g., EUPMS [80]), there was no reporting on feedback from developers and researchers who used these frameworks in various applications. Alongside the technical evaluations, reporting the measurable feedback from all the stakeholders will help others to identify which framework is suitable for them.

7.2.5 User Consent Support. When enrolling participants in a research study, user consent needs to be obtained. By giving consent, users give approval to be part of a study and for researchers to collect and process data for the specified purposes. Traditionally, such consent is given on paper by signing an informed consent document after having read a detailed description of the study or the purpose of the application. The introduction of new regulations such as General Data Protection Regulation (GDPR) [18] are empowering users to have full control over their data during and after the study is done, and puts forth the need for advanced consent management and configuration. ResearchKit [70] provides support for showing study information to the user on a phone and having him or her sign a consent form in a user-friendly manner. However, ResearchKit does not in any way support additional support for handling this consent during or after a study is done. Except for Bridge [11], all other frameworks in this review supporting the registration of consent leave further consent handling to the individual application that has been built using that framework. In addition, none of the frameworks provide support for handling partial consent or withdrawal of consent.

If user consent is not handled at the framework level, guaranteeing compliance with given consent during data sharing and processing within a framework may be challenging. A researcher might use data for another purpose than for which consent was given. Or, consent may expire, be modified, or retracted entirely. Manual tracking of consent forms and modifications made to them in a longitudinal study is challenging and negligence might pose privacy and legal risks. Providing support for obtaining and maintaining user consent in mHealth frameworks can help researchers identify which part of collected data the user has consented to and for which parts consent was withdrawn or modified. This would help in building more granular access restrictions for data processing and sharing.

7.2.6 Support for Study Reproducibility. When sensing frameworks are used to build applications for clinical and biomedical studies, support for study reproducibility becomes important [85]. The frameworks in this review offer limited or no support for study reproducibility. There have been some initiatives in this direction by Beiwe [85], which stores a configuration file specifying all of the applications settings to reproduce the study, such as surveys, user prompts, sensor settings, and version of the data analysis tool used. By standardizing data collection and data analysis, mHealth frameworks can help achieve study reproducibility and thereby encourage replication studies. For example, they could support exporting study protocols that define the study's technical know-how such as configuration and specifications of sensors and devices used, data sampling profile, defining when data should be collected, which stimuli need to be presented, tasks, interventions, and how user consent should be obtained for the experiment. The same study protocol could then be used by other researchers who wish to replicate the study.

7.2.7 Internationalization of Surveys. In software engineering, internationalization refers to tasks and activities by which a software system can adapt to different languages, cultural, and regional requirements without programmatic modifications (code) [93]. With the growing availability of smartphones, mHealth studies and interventions are reaching far beyond the English-speaking world. Therefore, it is essential for sensing frameworks to support various languages, text, and data input formats to reach all potential users. Only one framework (Ohmage [84]) in this review supports multiple languages for survey questions. However, it still does not support

full internationalization. Achieving full support for internationalization requires incorporating best practices such as string externalization, full support for Unicode, standard resource file types, and locale and culture awareness (e.g., date and time format, calendar differences) into the architecture of the sensing framework.

7.2.8 Gap between Collected mHealth Data and Clinical Knowledge. As discussed earlier, the reviewed frameworks focus primarily on supporting data collection. However, there is a big gap between collecting mHealth data and utilizing it according to clinical knowledge, such as incorporating it into workflow, decision-making, and so on, which has not been addressed in any of the reviewed frameworks. Previous studies have shown that in clinical settings getting actionable info from collected mHealth data remains a challenge due to limited time and resources [21, 89]. Therefore, future frameworks should also focus on bridging the gap between collecting information and obtaining actionable clinical knowledge from it.

7.3 Threats to Validity

There are two potential threats to validity that may affect our findings: the method adopted for *selecting articles* and *unavailable data* during data extraction.

7.3.1 Selecting Articles. As presented in Section 2, our review focuses on frameworks that may be used for health or behavior sensing. Some relevant articles may omit a discussion of the applicability of the presented framework in this context, in which case it would not show up in our search results. In addition, many of the data collection frameworks are published as part of a study. In such cases the main focus of the article is on the study and not the sensing framework. Thus, judging them solely on the basis of the title, it is likely that frameworks, which are described as part of a study article, might have been left out.

7.3.2 Unavailable Data. Since many articles do not explicitly report on all available functional and non-functional features, it is not always possible to assess whether or not a certain framework supports a given feature. Authors might omit mentioning certain features when they are already present in prior work, since they want to focus on novel contributions instead. Hence, we can only report on whether an article explicitly *describes* a particular feature. Therefore, features in the presented tables that do not include a check mark for a particular framework indicate the given feature was not discussed in the paper, but does not rule out entirely that it is not supported.

8 CONCLUSION

Generic sensing frameworks facilitate the development and deployment of mHealth applications and relieve developers and researchers from worrying about the underlying issues pertaining to smartphone and wearable sensor data collection, data offloading, processing, analysis, visualization, feature extraction, phone resource optimization, and study management. As the number of publications over time in this review indicate, research into the design and implementation of mHealth frameworks has grown in recent years. More recent frameworks include support for more and more functional and non-functional features. In this systematic literature review (SLR), we identified a list of 28 published and 9 unpublished open-source mHealth frameworks. Nearly one-third (N = 11) of the reviewed frameworks are still actively maintained.

We classified the functional features they support in three broad categories: sensing and storage, data processing and analysis, and study management. There are a total of 12 frameworks that support functional features in all three categories. With regard to non-functional features, our results suggest that there has not been enough emphasis on non-functional features in mHealth sensing frameworks and a large number of the articles just state supporting non-functional features without providing additional details such as necessary tradeoffs to make or a performance evaluation.

Based on this review, we discussed the state-of-the-art in mHealth sensing frameworks and identified new features that are relevant to explore in future research on such frameworks. We hope this overview of functional

and non-functional features, health studies in which existing frameworks have been used, how well frameworks have been maintained, and which license they use can help researchers and developers decide which framework is suitable for their next mHealth application.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers and the editor for their very valuable feedback on earlier drafts of the paper.

REFERENCES

[1] Kevin MacG Adams et al. 2015. *Nonfunctional Requirements in Systems Analysis and Design*. Vol. 28. Springer.
[2] Sasan Adibi. 2015. *Mobile Health: A Technology Road Map*. Vol. 5. Springer.
[3] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervas. Mob. Comput.* 7, 6 (2011), 643–659.
[4] Shahriar Akter and Pradeep Ray. 2010. mHealth—An ultimate platform to serve the unserved. *Yearb. Med. Inf.* 19, 01 (2010), 94–100.
[5] Talayeh Aledavood, Ana Maria Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Jari Saramäki, Erkki Isometsä, and Richard K. Darst. 2017. Data collection for mental health studies through digital platforms: Requirements and design of a prototype. *JMIR Res. Protoc.* 6, 6 (2017), e110.
[6] Apple. 2016. CareKit. Retrieved March 2, 2019 from http://carekit.org.
[7] Joost Asselbergs, Jeroen Ruwaard, Michal Ejdys, Niels Schrader, Marit Sijbrandij, and Heleen Riper. 2016. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: An explorative study. *J. Med. Internet Res.* 18, 3 (2016), e72.
[8] Sangwon Bae, Tammy Chung, Denzil Ferreira, Anind K. Dey, and Brian Suffoletto. 2018. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addict. Behav.* 83 (2018), 42–47.
[9] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proc. ACM Interact. Mob. Wearable Ubiq. Technol.* 1, 2 (June 2017). DOI: https://doi.org/10.1145/3090051
[10] Jakob E. Bardram and Mads Frost. 2016. The personal health technology design space. *IEEE Pervas. Comput.* 15, 2 (2016), 70–78.
[11] Sage Bionetworks. (2018). Bridge Sensing Platform. Retrieved March 2, 2019 from https://developer.sagebridge.org/articles/overview.html.
[12] Brian M. Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E. Ray Dorsey, et al. 2016. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci. Data* 3 (2016), 160011.
[13] Waylon Brunette, Rita Sodt, Rohit Chaudhri, Mayank Goel, Michael Falcone, Jaylen Van Orden, and Gaetano Borriello. 2012. Open data kit sensors: A sensor integration framework for Android at the application-level. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*. ACM, 351–364.
[14] Lora E. Burke, Jun Ma, Kristen M. J. Azar, Gary G. Bennett, Eric D. Peterson, Yaguang Zheng, William Riley, Janna Stephens, Svati H. Shah, Brian Suffoletto, et al. 2015. Current science on consumer use of mobile health for cardiovascular disease prevention: A scientific statement from the American Heart Association. *Circulation* 132, 12 (2015), 1157–1213.
[15] Rohit Chaudhri, Waylon Brunette, Bruce Hemingway, and Gaetano Borriello. 2013. ODK sensors: An application-level sensor framework for Android devices. In *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 30.
[16] Delphine Christin, Andreas Reinhardt, Salil S. Kanhere, and Matthias Hollick. 2011. A survey on privacy in mobile participatory sensing applications. *J. Syst. Softw.* 84, 11 (2011), 1928–1946.
[17] Chia-Fang Chung, Jonathan Cook, Elizabeth Bales, Jasmine Zia, and Sean A. Munson. 2015. More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. *J. Med. Internet Res.* 17, 8 (2015), e203.
[18] European Commission. 2019. Data protection. Retrieved March 2, 2019 from https://ec.europa.eu/info/law/law-topic/data-protection_en.
[19] IEEE Standards Coordinating Committee et al. 1990. IEEE standard glossary of software engineering terminology (IEEE Std 610.12-1990).. *IEEE Computer Society* Los Alamitos, CA.
[20] Confluent. 2019. *Kafka Connect | Confluent*. Retrieved February 28, 2019 from https://www.confluent.io/product/connectors/.
[21] Mary Jo Deering, Erin Siminerio, and Scott Weinstein. 2013. *Issue Brief: Patient-generated Health Data and Health IT. Office of the National Coordinator for Health Information Technology.*
[22] David Daniel Ebert, Pim Cuijpers, Ricardo F. Muñoz, and Harald Baumeister. 2017. Prevention of mental health disorders using internet- and mobile-based interventions: A narrative review and recommendations for future research. *Front. Psychiat.* 8 (2017), 116.

[23] Certification Europe. 2019. BS 10012 Personal Information Management System. Retrieved March 2, 2019 from https://www.certificationeurope.com/certification/bs-10012-personal-information-management-systems/.
[24] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: Mobile context instrumentation framework. *Front. ICT* 2 (2015), 6.
[25] Fitbit. 2015. Fitbit. Retrieved March 5, 2019 from https://www.fitbit.com.
[26] Funf. 2019. Funf. Retrieved March 2, 2019 from http://www.funf.org/about.html.
[27] Andrea Gaggioli, Giovanni Pioggia, Gennaro Tartarisco, Giovanni Baldus, Daniele Corda, Pietro Cipresso, and Giuseppe Riva. 2013. A mobile data collection platform for mental health research. *Pers. Ubiq. Comput.* 17, 2 (Feb. 2013), 241–251. DOI: https://doi.org/10.1007/s00779-011-0465-2
[28] GoogleFit. 2014. Google Fit. Retrieved April 2, 2019 from https://www.google.com/fit/.
[29] Mohammad Hashemian, Dylan Knowles, Jonathan Calver, Weicheng Qian, Michael C. Bullock, Scott Bell, Regan L. Mandryk, Nathaniel Osgood, and Kevin G. Stanley. 2012. iEpi: An end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In *Proceedings of the 2nd ACM International Workshop on Pervasive Wireless Healthcare*. ACM, 3–8.
[30] HealthKit. 2014. HealthKit. Retrieved April 2, 2019 from https://developer.apple.com/healthkit/.
[31] Netzahualcóyotl Hernández, Bert Arnrich, Jesús Favela, Cem Ersoy, Burcu Demiray, and Jesús Fontecha. 2019. A multi-site study on walkability, data sharing and privacy perception using mobile sensing data gathered from the mk-sense platform. *J. Amb. Intell. Hum. Comput.* 10, 6 (2019), 2199–2211.
[32] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen, and Deborah Estrin. 2010. AndWellness: An open mobile system for activity and experience sampling. In *Proceedings of the Wireless Health Conference*. ACM, 34–43.
[33] Mark D. Hill. 1990. What is scalability?*ACM SIGARCH Comput. Archit. News* 18, 4 (1990), 18–21.
[34] Health Level 7 (HL7). 2016. HL7 Mobile Health. Retrieved December 2, 2019 from http://www.hl7.org/Special/committees/mobile/index.cfm.
[35] Syed Monowar Hossain, Timothy Hnat, Nazir Saleheen, Nusrat Jahan Nasrin, Joseph Noor, Bo-Jhang Ho, Tyson Condie, Mani Srivastava, and Santosh Kumar. 2017. mCerebrum: A mobile sensing software platform for development and validation of digital biomarkers and interventions. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 7.
[36] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 493–504.
[37] Cheng-Kang Hsieh, Hongsuda Tangmunarunkit, Faisal Alquaddoomi, John Jenkins, Jinha Kang, Cameron Ketcham, Brent Longstaff, Joshua Selsky, Betta Dawson, Dallas Swendeman, et al. 2013. Lifestreams: A modular sense-making toolset for identifying important patterns from everyday life. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 5.
[38] Intel. 2015. Intel Context Sensing SDK. Retrieved March 12, 2019 from https://software.intel.com/en-us/context-sensing-sdk/features.
[39] Kleomenis Katevas, Hamed Haddadi, and Laurissa Tokarchuk. 2014. Poster: Sensingkit: A multi-platform mobile sensing framework for large-scale experiments. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*. ACM, 375–378.
[40] Alexander D. Kent and Lorie M. Liebrock. 2011. Secure communication via shared knowledge and a salted hash in ad-hoc environments. In *Proceedings of the IEEE 35th Annual Computer Software and Applications Conference Workshops*. IEEE, 122–127.
[41] Jayden Khakurel, Helinä Melkas, and Jari Porras. 2018. Tapping into the wearable device revolution in the work environment: A systematic review. *Inf. Technol. People* 31, 3 (2018), 791–818.
[42] Wazir Zada Khan, Yang Xiang, Mohammed Y. Aalsalem, and Quratulain Arshad. 2012. Mobile phone sensing systems: A survey. *IEEE Commun. Surv. Tutor.* 15, 1 (2012), 402–427.
[43] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01. Software Engineering Group, School of Computer Science and Mathematics, Keele University and Department of Computer Science, University of Durham.
[44] Dylan L. Knowles, Kevin G. Stanley, and Nathaniel D. Osgood. 2014. A field-validated architecture for the collection of health-relevant behavioural data. In *Proceedings of the IEEE International Conference on Healthcare Informatics*. IEEE, 79–88.
[45] Youngki Lee, S. S. Iyengar, Chulhong Min, Younghyun Ju, Taiwoo Park, Jinwon Lee, Yunseok Rhee, and Junehwa Song. 2012. Mobicon: A mobile context-monitoring platform. *Commun. ACM* 55, 3 (2012), 54.
[46] Young-Hee Lee and Ryang-Hee Kim. 2018. Estimation of the smartphone user—Satisfaction and customer intention on the social networking service. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*. Springer, 262–271.
[47] Xinyi Li, Elizabeth Vera, Mark Gilbert, Orieta Celiku, and Terri Armstrong. 2018. INNV-41. My STORI—a symptom tracking and reporting instrument mobile application for central nervous system cancer patients.*Neuro-Oncol.* 20 (11 2018), vi146–vi146. DOI: https://doi.org/10.1093/neuonc/noy148.609
[48] Jong Hyun Lim, Andong Zhan, Evan Goldschmidt, JeongGil Ko, Marcus Chang, and Andreas Terzis. 2012. HealthOS: A platform for pervasive health applications. In *Proceedings of the 2nd ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*. ACM, 4.

[49] Felix Xiaozhu Lin, Ahmad Rahmati, and Lin Zhong. 2010. Dandelion: A framework for transparently programming phone-centered wireless body sensor applications for health. In *Proceedings of the Wireless Health Conference (WH'10)*. ACM, New York, NY, 74–83. DOI: https://doi.org/10.1145/1921081.1921091

[50] Carissa A. Low, Anind K. Dey, Denzil Ferreira, Thomas Kamarck, Weijing Sun, Sangwon Bae, and Afsaneh Doryab. 2017. Estimation of symptom severity during chemotherapy from passively sensed data: Exploratory study. *J. Med. Internet Res.* 19, 12 (19 Dec. 2017), e420. DOI: https://doi.org/10.2196/jmir.9046

[51] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2010. The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 71–84.

[52] Rischan Mafrur, I. Gde Dharma Nugraha, and Deokjai Choi. 2015. Modeling and discovering human behavior from smartphone sensing life-log data for identification purpose. *Hum.-cent. Comput. Inf. Sci.* 5, 1 (12 Oct. 2015), 31. DOI: https://doi.org/10.1186/s13673-015-0049-7

[53] Alex Mariakakis, Edward Wang, Shwetak Patel, and Mayank Goel. 2019. Challenges in realizing smartphone-based health sensing. *IEEE Pervas. Comput.* 18, 2 (2019), 76–84.

[54] Wendy Berry Mendes. 2018. My BP Lab. Retrieved March 2, 2019 from http://sagebionetworks.org/research-projects/my-bp-lab/.

[55] Open mHealth. 2015. Case study: Post-Traumatic Stress (PTSD). Retrieved March 2, 2019 from http://www.openmhealth.org/features/case-studies/case-study-post-traumatic-stress-ptsd/.

[56] Open mHealth. 2015. Case study: Type 1 diabetes. Retrieved March 2, 2019 from http://www.openmhealth.org/features/case-studies/case-study-type-1-diabetes/.

[57] Open mHealth. (2019). Open mHealth. Retrieved March 5, 2019 from http://www.openmhealth.org/.

[58] Chulhong Min, Chungkuk Yoo, Youngki Lee, and Junehwa Song. 2011. Healthopia: Towards your well-being in everyday life. In *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies*. ACM, 108.

[59] Xiaoyun Mo, Dianxi Shi, Ruosong Yang, Han Li, ZheHang Tong, and Feng Wang. 2015. A framework of fine-grained mobile sensing data collection and behavior analysis in an energy-configurable way. In *Proceedings of the IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity'15)*. IEEE, 391–398.

[60] W. Morrison, L. Guerdan, J. Kanugo, T. Trull, and Y. Shang. 2018. TigerAware: An innovative mobile survey and sensor data collection and analytics system. In *Proceedings of the IEEE 3rd International Conference on Data Science in Cyberspace (DSC'18)*. 115–122. DOI: https://doi.org/10.1109/DSC.2018.00025

[61] Institute of Electrical and Electronics Engineers. 2019. IEEE 11073 Personal health devices. Retrieved December 2, 2019 from https://standards.ieee.org/standard/11073-00103-2012.html.

[62] Institute of Electrical and Electronics Engineers. 2019. Standard for Mobile Health Data. Retrieved December 2, 2019 from https://standards.ieee.org/project/1752.html.

[63] National Institute of Standards and Technology. 2001. Advanced Encryption Standard (AES). Retrieved March 2, 2019 from https://www.nist.gov/publications/advanced-encryption-standard-aes.

[64] Charith Perera, Prem Prakash Jayaraman, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2014. MOSDEN: An internet of things middleware for resource constrained mobile devices. In *Proceedings of the 47th Hawaii International Conference on System Sciences*. IEEE, 1053–1062.

[65] Pascal B. Pfiffner, Isaac Pinyol, Marc D. Natter, and Kenneth D. Mandl. 2016. C3-PRO: Connecting ResearchKit to the health system using i2b2 and FHIR. *PloS One* 11, 3 (2016), e0152722.

[66] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. 2010. EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 281–290.

[67] Satish Ramkumar, Nitesh Nerlekar, Daniel D'Souza, Derek J. Pol, Jonathan M. Kalman, and Thomas H. Marwick. 2018. Atrial fibrillation detection using single lead portable electrocardiographic monitoring: A systematic review and meta-analysis. *BMJ Open* 8, 9 (2018), e024178.

[68] Yatharth Ranjan, Maximilian Kerz, Zulqarnain Rashid, Sebastian Böttcher, Richard J. B. Dobson, and Amos A. Folarin. 2018. RADAR-base: A novel open source m-health platform. In *Proceedings of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 223–226.

[69] Reza Rawassizadeh, Martin Tomitsch, Katarzyna Wac, and A. Min Tjoa. 2013. UbiqLog: A generic mobile phone-based life-log framework. *Person. Ubiq. Comput.* 17, 4 (2013), 621–637.

[70] ResearchKit. 2015. ResearchKit. Retrieved March 2, 2019 from http://researchkit.org/.

[71] ResearchStack. 2016. ResearchStack. Retrieved March 5, 2019 from http://researchstack.org/.

[72] Purple Robot. 2015. Purple Robot. Retrieved March 2, 2019 from https://tech.cbits.northwestern.edu/purple-robot/.

[73] Darius A. Rohani, Nanna Tuxen, Lars V. Kessing, and Jakob E. Bardram. 2017. Designing for hourly activity sampling in behavioral activation. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'17)*. ACM, New York, NY, 431–435. DOI: https://doi.org/10.1145/3154862.3154919.

[74] Ángel Ruiz-Zafra, Kawtar Benghazi, Manuel Noguera, and José Luis Garrido. 2013. Zappa: An open mobile platform to build cloud-based m-health systems. In *Ambient Intelligence-software and Applications*. Springer, 87–94.

[75] Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *J. Med. Internet Res.* 17, 7 (2015), e175.

[76] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. 2015. puffMarker: A multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 999–1010.

[77] Akio Sashima, Yutaka Inoue, Takeshi Ikeda, Tomohisa Yamashita, and Koichi Kurumatani. 2008. CONSORTS-S: A mobile sensing platform for context-aware services. In *Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 417–422.

[78] Akio Sashima, Yutaka Inoue, Takeshi Ikeda, Tomohisa Yamashita, Masayuki Ohta, and Koichi Kurumatani. 2008. Toward mobile healthcare services by using everyday mobile phones. In *Proceedings of the 1st International Conference on Health Informatics*. 242–245.

[79] M. Schickler, R. Pryss, M. Stach, J. Schobel, W. Schlee, T. Probst, B. Langguth, and M. Reichert. 2017. An IT platform enabling remote therapeutic interventions. In *Proceedings of the IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS'17)*. 111–116. DOI: https://doi.org/10.1109/CBMS.2017.78.

[80] Johannes Schobel, Rüdiger Pryss, Marc Schickler, Martina Ruf-Leuschner, Thomas Elbert, and Manfred Reichert. 2016. End-user programming of mobile services: Empowering domain experts to implement mobile data collection applications. In *Proceedings of the IEEE International Conference on Mobile Services (MS'16)*. IEEE, 1–8.

[81] Johannes Schobel, Rüdiger Pryss, Wolfgang Wipp, Marc Schickler, and Manfred Reichert. 2016. A mobile service engine enabling complex data collection applications. In *Proceedings of the International Conference on Service-oriented Computing*. Springer, 626–633.

[82] Audacious Software. 2016. Passive Data Kit. Retrieved March 2, 2019 from https://audacious-software.com/passive-data-kit/.

[83] Callum L. Stewart, Zulqarnain Rashid, Yatharth Ranjan, Shaoxiong Sun, Richard J. B. Dobson, and Amos A. Folarin. 2018. RADAR-base: Major depressive disorder and epilepsy case studies. In *Proceedings of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 1735–1743.

[84] Hongsuda Tangmunarunkit, Cheng-Kang Hsieh, Brent Longstaff, S. Nolen, John Jenkins, Cameron Ketcham, Joshua Selsky, Faisal Alquaddoomi, Dony George, Jinha Kang, et al. 2015. Ohmage: A general and extensible end-to-end participatory sensing platform. *ACM Trans. Intell. Syst. Technol.* 6, 3 (2015), 38.

[85] John Torous, Mathew V. Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Ment. Health* 3, 2 (2016).

[86] Jiangtao Wang, Yasha Wang, Daqing Zhang, and Sumi Helal. 2018. Energy saving techniques in mobile crowd sensing: Current state and future opportunities. *IEEE Commun. Mag.* 56, 5 (2018), 164–169.

[87] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[88] Dan E. Webster, Christine Suver, Megan Doerr, Erin Mounts, Lisa Domenico, Tracy Petrie, Sancy A. Leachman, Andrew D. Trister, and Brian M. Bot. 2017. The mole mapper study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. *Sci. Data* 4 (2017), 170005.

[89] Peter West, Max Van Kleek, Richard Giordano, Mark J. Weal, and Nigel Shadbolt. 2018. Common barriers to the use of patient-generated data across clinical settings. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 484.

[90] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 38.

[91] Pang Wu, Huan-Kai Peng, Jiang Zhu, and Ying Zhang. 2011. Senscare: Semi-automatic activity summarization system for elderly care. In *Proceedings of the International Conference on Mobile Computing, Applications, and Services*. Springer, 1–19.

[92] Pang Wu, Jiang Zhu, and Joy Ying Zhang. 2013. MobiSens: A versatile mobile sensing platform for real-world applications. *Mob. Netw. Applic.* 18, 1 (2013), 60–80.

[93] X. Xia, D. Lo, F. Zhu, X. Wang, and B. Zhou. 2013. Software internationalization and localization: An industrial experience. In *Proceedings of the 18th International Conference on Engineering of Complex Computer Systems*. 222–231. DOI: https://doi.org/10.1109/ICECCS.2013.40.

[94] Haoyi Xiong, Yu Huang, Laura E. Barnes, and Matthew S. Gerber. 2016. Sensus: A cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 415–426.

Received December 2019; revised May 2020; accepted July 2020

A.2 mCardia: An ambulatory Context-Aware ECG Collection System for Arrhythmia Screening

Authors
Devender Kumar, Rjau Mahrajan, Alban Maxhuni, Helena Dominguez, Anne Frølich & Jakob Eyvind Bardram

Journal
ACM Transactions on Computing for Healthcare [Accepted]

-

mCardia: An Ambulatory Context-Aware ECG Collection System for Arrhythmia Screening

DEVENDER KUMAR, Department of Health Technology, Technical University of Denmark, Denmark
RAJU MAHARJAN, Department of Health Technology, Technical University of Denmark, Denmark
ALBAN MAXHUNI, Department of Health Technology, Technical University of Denmark, Denmark
HELENA DOMINGUEZ, Bispebjerg-Frederiksberg Hospital, Department of Cardiology, Denmark
ANNE FRØLICH, Department of Public Health, University of Copenhagen, Denmark
JAKOB E. BARDRAM, Department of Health Technology, Technical University of Denmark, Denmark

This paper presents the design, development, and evaluation of the usability and feasibility of *mCardia* – a context-aware, mobile electrocardiogram (ECG) collection system for longitudinal arrhythmia screening under free-living conditions. Along with ECG, *mCardia* also records active and passive context data, including patient-reported symptoms and physical activity. This contextual data can provide a more accurate understanding of what happens before, during, and after an arrhythmia event, thereby providing additional information in the diagnosis of arrhythmia. By using a plugin-based architecture for ECG and contextual sensing, *mCardia* is device-agnostic and can integrate with various wireless ECG devices and support cross-platform deployment. We deployed the *mCardia* system in a two-week feasibility study involving 24 patients. During the study, we observed high patient acceptance and compliance with a satisfactory yield of collected ECG and contextual data. The results demonstrate high usability and feasibility of *mCardia* for longitudinal ambulatory monitoring under free-living conditions. The paper also reports from two clinical cases, which demonstrates how a cardiologist can utilize the contextual data collected to improve the accuracy of arrhythmia analysis. Finally, the paper discusses the lessons learned and the challenges found in the *mCardia* design and feasibility study.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Software and its engineering → Development frameworks and environments**; • **Applied computing →** Health informatics.

Additional Key Words and Phrases: arrhythmia, arrhythmia screening, mobile health, mHealth, ECG, mobile sensing, Holter monitoring, cardiovascular, context aware ECG

ACM Reference Format:
Devender Kumar, Raju Maharjan, Alban Maxhuni, Helena Dominguez, Anne Frølich, and Jakob E. Bardram. 2021. mCardia: An Ambulatory Context-Aware ECG Collection System for Arrhythmia Screening . *J. ACM* -, -, Article - (2021), 28 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Devender Kumar, Department of Health Technology, Technical University of Denmark, Lyngby,2800, Copenhagen , Denmark, deku@dtu.dk; Raju Maharjan, Department of Health Technology, Technical University of Denmark, Lyngby,2800, Copenhagen , Denmark, rajm@dtu.dk; Alban Maxhuni, Department of Health Technology, Technical University of Denmark, Lyngby,2800, Copenhagen , Denmark, almax@dtu.dk; Helena Dominguez, Bispebjerg-Frederiksberg Hospital, Department of Cardiology, Lyngby,2800, Copenhagen, Denmark, maria.helena.dominguez.vall-lamora.02@regionh.dk; Anne Frølich, Department of Public Health, University of Copenhagen, Bispebjerg,2450, Copenhagen , Denmark, anfro@sund.ku.dk; Jakob E. Bardram, Department of Health Technology, Technical University of Denmark, Lyngby,2800, Copenhagen , Denmark, jakba@dtu.dk.

0004-5411/2021/–ART- $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Cardiovascular diseases (CVD) are the most prevailing diseases globally causing 17.9 million deaths (nearly 31% of all global deaths) in the year 2016 alone [48]. Amongst the CVD, cardiac arrhythmias are widespread and affect over four million people in the US, costing up to $67.7 billion annually [56]. Cardiac arrhythmias such as Atrial Fibrillation (AF) and Atrial Flutter are the most common causes of heart failure, hospitalization, thromboembolic events [10, 13] and death. Nearly 30% of patients with AF are unaware of their diagnosis [18]. Therefore, early diagnosis and treatment of such arrhythmias is important to provide timely healthcare and prevent life threatening conditions [25]. Moreover, CVD are chronic diseases, and the provision of chronic care in CVD patients does not reach quality standards in about 50% of cases, which has significant consequences for both patients and society [43]. To meet the demand for high-quality care based on evidence-based principles, organizational care models such as the Chronic Care Model (CCM) [60] have been proposed. Ambulatory technology for continuous CVD monitoring is seen as an essential means for improving chronic care outcome [21].

1.1 Traditional arrhythmia screen process

AF and other heart rhythm disorders are diagnosed based on ECG analysis. The current practice of arrhythmias diagnosis employs a 6 or 4-channel Holter monitor[1] and records ECG data from the patients, typically for 24-48 hrs. During this period, the patient is instructed to maintain a diary for keeping track of any symptoms suspected to be related to arrhythmia (e.g., palpitation, dizziness, chest pain). Figure 1 shows the typical workflow of the traditional home-based ECG Holter monitoring. When the recording is complete, the patient-reported diary and ECG recordings from the Holter monitor are examined by a trained Holter nurse or a cardiologist. This method of Holter monitoring is subjected to many limitations. For instance, when arrhythmias are often sporadic, it may not be easy to detect them in just 24-48 hours of Holter monitoring. Long-term ambulatory ECG under the free-living condition is limited by the Holter device's battery life and the need for turn-over of the devices to examine other patients. Even with seven-day Holter monitoring, about 30 percent of episodes are missed [32]. Implanted pacemakers provide an alternative way of collecting long term ECG for arrhythmia analysis, but they are invasive. Similarly, implantable loop-recorders can also sense arrhythmias for very prolonged periods (years). However, compared to continuous Holter monitoring, their capacity to effectively detect arrhythmia is limited due to their dependence on their algorithm and limited memory capacity [19, 26, 59]. Recent advances in wearable and mobile health (mHealth) technologies, however, have made it possible to collect long term ECG in a non-invasive way, and several studies [22, 44] have shown that they are also very cost-effective for arrhythmia diagnosis and management.

1.2 Challenges in ambulatory ECG monitoring

Although longitudinal ambulatory ECG monitoring under natural settings using wired Holter monitors are useful for improving arrhythmia screening, it has a set of challenges.

1.2.1 Arrhythmia misinterpretation due to the lack of contextual information. When recording long-term ECG under free-living conditions, many motion artifacts and anomalies are added to the ECG signal. These artifacts mimic arrhythmias [42], which impacts both manual and computer-aided automated arrhythmia diagnostic accuracy, resulting in more false positive and false negative detections. A recent smartwatch-based arrhythmia detection study conducted by Dörr et al. [15]

[1]ECG monitors for ambulatory use is often called a 'Holter' monitor in general. This name comes from the Holter Research Laboratory where experimental physicists Norman J. Holter and Bill Glasscock started working on radio telemetry in 1949.

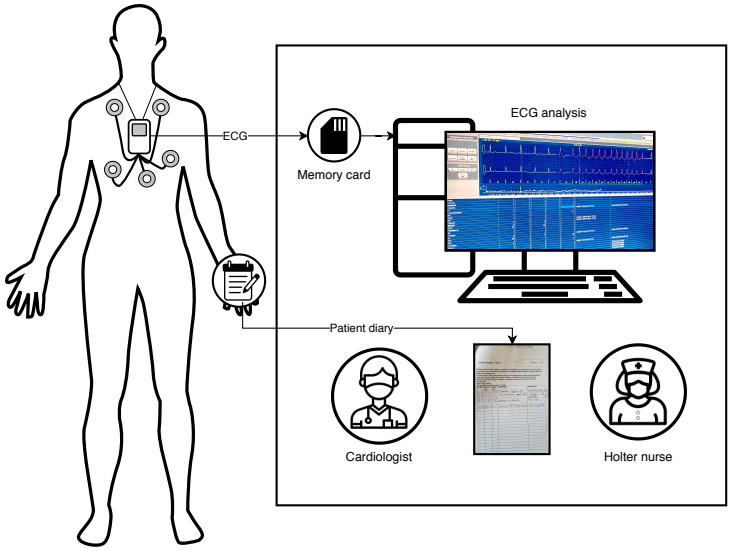


Fig. 1. Workflow of the traditional outpatient ECG Holter monitoring during arrhythmia screening. The patient wears a four or six-lead wired ECG monitor for 24 to 72 hours and fills out a diary with the symptoms experienced during this period. When the patient returns to the outpatient clinic, the ECG data is transferred (using a USB stick) to a large desktop computer where a Holter nurse or cardiologist analyzes it. During the ECG analysis, the Holter nurse and the cardiologist use the patient diary to understand the ECG recording in comparison to all reported symptoms.

reported that factors such as movement artifacts and poor signal quality adversely affected the diagnostic performance of arrhythmia detection algorithms when applied to ambulatory ECG.

Moreover, the lack of knowledge about the context in which the ECG was recorded can lead to misinterpretation and misclassification of heart arrhythmias [2, 11]. Dinakarrao et al. [14] recently reviewed the trends and techniques of computer-aided arrhythmia diagnosis and highlighted that ECG analysis independent of the patients' physical condition and context (such as activities, place, food intake) could induce misinterpretation and misclassification of arrhythmias, which is likely to multiply manifold when dealing with long term ECG recordings from the free-living condition, due to a number of confounding artifacts.

1.2.2 Lack of user engagement in longitudinal ECG collection. In longitudinal home-based ECG monitoring for arrhythmia screening, patients need to be active participants. However, previous research have shown that sustained patient engagement is a major barrier in collecting quality ambulatory ECG [51]. In addition, the use of traditional wired Holter monitors for longitudinal ECG collection under free-living conditions is very inefficient as they are bulky in size [31] and do not provide feedback for patient engagement during the process, and hence remains a 'black box' to the patient.

1.2.3 Poor signal quality. In long-term ambulatory ECG collection, it is quite common that electrodes become non-adhesive over time resulting in poor ECG signals. While some Holter monitors provide alarms to the patient on poor signal quality, patients are not always able to identify and take corrective measures to fix or change the electrodes, which results in noisy or unusable ECG

data. Additionally, the motion artifacts during various activities under free-living conditions affects the signal quality [49].

1.2.4 Recall bias on patient's self-reported symptoms and events. The quality, usefulness, and reliability of patient-generated data without a full understanding of the patient's context have been a major concern in clinical settings [61]. As shown in Figure 1, during ambulatory Holter monitoring, patients are asked to keep a diary of any cardio-related symptoms or events (e.g., dizziness, palpation, shortness of breathing). The diary is typically paper-based. The cardiologists use the information in this diary during ECG analysis by mapping symptoms noted in the diary with the ECG recording, in order to find any occurrence of arrhythmias. This paper-based diary suffers from recall bias as patients do not always carry it with them and have to rely on their memory when filling in the diary later. Moreover, the timing noted in the diary is often wrong or very imprecise. This lack of synchronization between the timestamp of the patient-reported symptoms with the ECG signal's timestamp makes it difficult for the cardiologist or nurse to map the symptoms to the corresponding ECG during analysis. This recall bias on patient-reported symptoms gets multiplied manifold when collecting long term ECG in free-living conditions.

1.3 Context-aware ECG monitoring under free-living conditions

To address the above-mentioned challenges in longitudinal arrhythmia screening, this paper presents the design, implementation, and usability and feasibility evaluation of *mCardia*, which is a context-aware ECG collection system designed to be used under free-living conditions. In contrast to conventional ECG monitoring, *mCardia* provides means to collect contextual information such as activities, sleep quantity and quality, body position (sleeping left/right/supine), user-provided abnormal symptoms and events (e.g., dizziness, palpitation) and their duration, food intake, and maps them with the raw ECG recordings (see Figure 2). This contextual information can help cardiologists reducing misinterpretation and misclassification of events due to many confounding artifacts when assessing long-term ECG data for arrhythmia screening. In addition, these contextual features associated with the raw ECG data can be utilized as input to machine learning algorithms for reducing the false alarm in automatic arrhythmia analysis of ambulatory ECG. In addition, *mCardia* can also help in improving doctor and patient communication in a chronic care model.

The contribution of this work is in (1) identifying relevant contextual data that can help in improving arrhythmia diagnosis, (2) implementation of *mCardia* system based on a device-independent plugin-based software architecture which makes it easy to integrate with other ECG devices, and (3) demonstrating the usability and feasibility of such a system in longitudinal arrhythmia screening via field deployment and two case studies.

The remainder of this paper consists of 8 sections. Section 2 describes related work. Section 3 outlines the design process of the proposed system and section 4 describes the *mCardia* system's architecture and implementation. Sections 5 and 6 describe the usability and clinical feasibility study and its results. Section 7 discusses the obtained results and section 8 lists the limitations. Thereafter, section 9 concludes the paper.

2 RELATED WORK

The long term in situ ECG monitoring is an area that is rapidly growing and very promising for early diagnosis and follow-up of Cardiovascular diseases. In particular, ECG monitoring for arrhythmia detection has received a great deal of attention with a plethora of remote ECG monitoring devices such as patches, wired and wireless Holters, and event recorders [53]. Event recorders are the devices that record an ECG after or during an arrhythmia event. They record ECG for a short duration (30 sec to 1 min) and transmit it to a clinician or a heart monitoring center. These events are

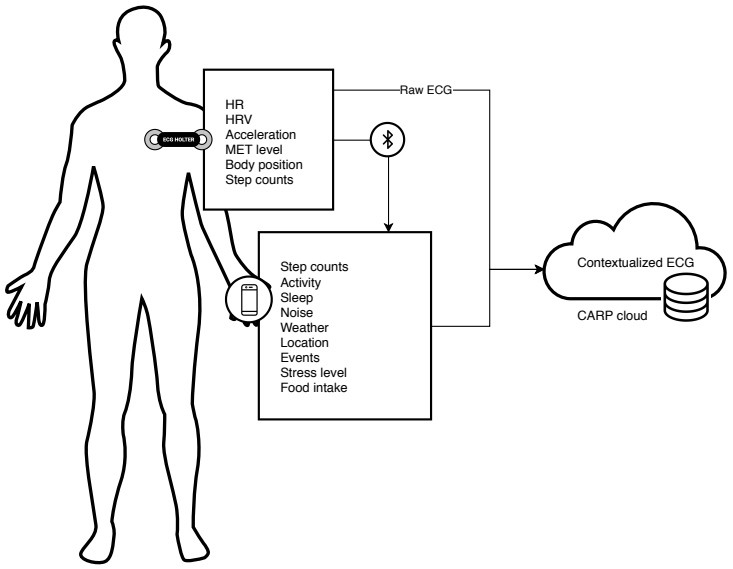


Fig. 2. Schematic diagram of *mCardia*. The ECG device collects raw ECG in combination with a set of other biomedical signals (HR, HRV, etc.). Contextual data is collected from the phone. All data is transmitted from the phone to a cloud-based data management system.

either marked by patients or, in some cases, by the recorder's auto-activation algorithm. Compared to the event recorders, 'patch monitors' and 'Holter Monitors' provide ECG collection for a longer duration (2-7 days). They operate in both modes: (1) real-time transmission, (2) store data for later analysis. Patches are more comfortable than Holters and mostly wireless [53]. However, most of the above-mentioned ECG collection systems only collect ECG and do not capture the user's context in which the ECG was collected.

2.1 Contextualized Mobile ECG

Some of the earliest work on contextualized mobile ECG was done by Shirazi et al. [52] in CardioViz, where they showed the ECG signal together with photos and notes overlaid on a map. The core idea of the system was to use a phone camera for photography and an external GPS sensor, that can hint and remind users about the place and location related to a specific period in ECG. Similarly, Belgacem and Boumerdassi [9] proposed a mobile ECG monitoring system with the patient location. The contextual information collected in these two systems was limited to GPS location and photography. Moreover, neither of these systems were evaluated for their usability and feasibility in a clinical setting.

Forkan et al. [16] proposed context-aware cardiac monitoring systems for reducing false alerts by utilizing the context-awareness of the collected ECG data. In their other work [17] they developed a context-aware system for monitoring elder cardiac patients in the home environment and integrated it with social networking services for distant help and tracking, by shared the contextual health situation with friends, family, and doctors. Their focus was on expanding the elderly patient's social linkage and not on the effective use of contextual data for improving arrhythmia diagnosis itself.

Spadini et al. [54] proposed the iMote2-based context-aware ECG monitoring system, which, along with ECG, also collects accelerometer and environmental data such as temperature, humidity, and light intensity. Although this system collected some passive context from sensors, it does not

take subjective, or user's self-described context/symptoms into consideration. Contrary to this, the context-aware ECG monitoring system proposed by Li et al. [39] mainly focused on active context collection, which was also just limited to activity recognition. Upon detecting any change in the bio-signal recording system, it would prompt interface to the user, asking them to enter the activity they are performing. Similarly, Miao et al. [46] built a wearable Context-Aware ECG Monitoring by utilized built-in smartphone sensors for physical activity recognition. They demonstrated that physical activity recognition could help improve the diagnosis accuracy of heart arrhythmias and identify the most frequent irregular ECG patterns in various activities. In line with the finding from the Miao et al. [46]'s work, *mCardia* is also designed to utilize the built-in kinematic sensors of smartphone for collecting user context. However, compared to their work where context collection was limited to physical activity recognition, *mCardia* also collects user reported (subjective) context, which is very essential when analyzing long term ECG for arrhythmia detection. Furthermore, they did not evaluate the usability of the proposed system under free-living conditions, and the feedback provided to the user during the data collection was also limited.

2.2 Commercial patches and short term ECG measurements apps

In addition, there are commercial patches such as Zio Patch [29], WebCardio [40] which are designed for short term (1-2 days) to long term (over 14 days) ECG collection. However, they also do not collect contextual information, and patients are asked to maintain a symptoms/events diary and send it to the doctor together with the device after the screening period.

In contrast to *mCardia* which focuses on longitudinal continuous ECG collection, in recent years there has been may mobile ECG /PPG measurement apps and devices such as AliveCor [1], Apple watch [28], CardiioRhythm [12], and Zenicor-ECG [55] are being utilized for short term (30 sec – 1 min) ECG recording for arrhythmia screening. In this type of ECG measurement, patients usually take short ECG recording several times a day or when they feel any unusual symptoms and share it with the doctor. Nevertheless, in many cases, continuous long-term monitoring is desirable as short term sporadic ECG might miss asymptomatic arrhythmias [20].

3 RESEARCH METHODS

The design of the *mCardia* system followed a User-centered Design (UCD) approach [23] and applied the patient-clinician-designer (PCD) framework [41]. This research method seeks to find a good design compromise by considering different, and sometimes conflicting, concerns from the perspective of three stakeholder groups; the patient, the clinician, and the designer. The PCD framework provides a structured process for mediating co-design activities to find appropriate design solutions. In total, the design activities spanned 16 months (April 2018 to July 2019) and involved 14 participants; 6 patients, 4 clinicians (3 cardiologists (MD) and 1 nurse), and 4 designers.

The entire design process involved clinicians affiliated with a department of cardiology at a high-volume Danish University Hospital, in Copenhagen. Patients were also recruited from this department. Figure 3 illustrates the timeline of the design process and its main activities, which involved identifying the context of use, specifying requirements, and two iterations of design and evaluation.

3.1 Identify the context of use

In this phase, the design involved a group of clinicians including a cardiologist, a nurse, and a medical intern in the department of cardiology, and a group of three patients.

To identify the clinicians' needs and context of use of the system, we conducted several design meetings and workshops focusing on designing three main parts of the system: (a) What data to collect, including cardiovascular data (e.g., ECG and Heart Rate (HR)), contextual data (e.g.,

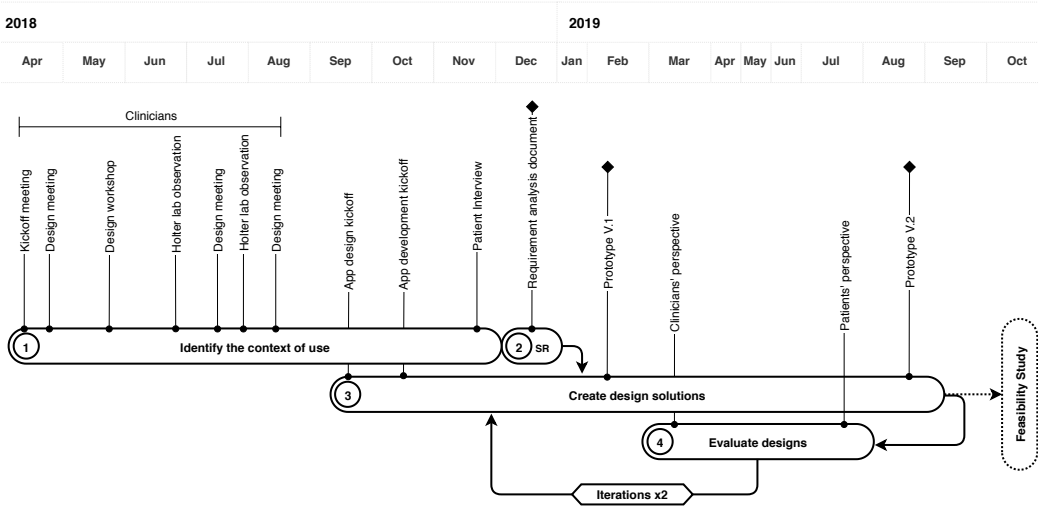


Fig. 3. User-centered Design (UCD) process timeline. SR = Specify Requirements

physical activity and step count), self-reported symptoms (e.g., feeling dizzy or experiencing racing heartbeats), and self-reported lifestyle data (e.g., information on sleep and eating). (b) The user experience, including creating the User Experience (UX) design for how to use and mount the ECG device, how to pair it with the phone, how to input data on the phone, how data is visualized, and how to navigate between different parts of the app. (c) Using the data in clinical practice, including understanding how the contextual and self-reported data can improve arrhythmia screening and diagnosis and how data can be mapped and visualized. Meeting minutes were noted and circulated to the participants from each meetings.

In parallel to the design workshops, we conducted an observational field study of the current Holter monitoring process (Figure 1) in the hospital's outpatient clinic. This study focused on observing and understanding the current process of preparing and setting up a patient for Holter monitoring, including how the device was introduced and mounted on patients, as well as getting the device and data back from the patient. The study also observed the process of analyzing the ECG data and the paper-based diaries collected from the patients, and how they were analyzed in the current software systems. Detailed notes were taken throughout this observational study. Pictures were taken and the audio was recorded while the nurse analyzed the ECG data following a think-aloud protocol.

To understand the patients' needs and context of use of the system, we involved and interviewed three patients (P1: M/78, P2: F/73, P3: F/25), who had been subject to traditional Holter monitoring. The patients were involved in the study outside the cardiology department, when they finished their Holter monitoring, and returned to the outpatient clinic. Open-ended interviews were conducted that focused on their experience using the traditional Holter monitoring setup. Notes were taken during the interviews.

3.2 Requirements Specification

Based on the first phase of the project, a comprehensive list of requirements were collected and documented in a Requirement Analysis Document (RAD) [36]. In summary, these requirements are:

- *User engagement in data collection.* In terms of the user experience, the primary concern of both patients and clinicians was the collection and mapping of self-reported symptoms and activities. Traditional Holter monitoring typically involves collection of self-reported symptoms and activities on paper-based diaries filled in by the patient. Later the clinicians need to map this to the ECG recordings. Clinicians stated that this process was cumbersome and the data was not valid, since the paper diaries suffered from many flaws including missing data, incomplete data, and recall bias. All the patients agreed that it would be much easier for them to record the symptoms and activities in a smartphone app. For example, P1 only entered activities such as cycling, running etc. at the end of day and often forgot this, and P2 did not record anything in the paper diary. P3 explained that she took notes on her smartphone and then had to manually add this to the paper diary at the end of the day. Patients also reported that it was difficult to push the event marker button on the Holter device, since it was hard to locate underneath clothes. Therefore, the requirement was to provide better methods for collecting self-reported symptoms and activities, and automatically map these with the timeline of the ECG recording.
- *Collect contextual information relevant for arrhythmia screening.* Clinicians emphasized the value of contextual information such as HR, Heart Rate Variability (HRV), sleep, step counts, Metabolic (MET) level, self-reported symptoms, symptom duration, and patients' activity during the symptom in order to better interpret the ECG data. Patients were willing to provide these information as long as they are informed about what data is being collected and how it is used for the diagnosis and treatment purposes. Therefore, the requirement was to collect a wide set of contextual information about the patient's conditions, activities and where-abouts, by implementing ecologically momentary assessment (EMA) approach as well as by collecting data automatically from the sensors on the ECG device and the smartphone.
- *System feedback.* The clinicians opposed the idea of providing visual feedback on physiological data to the patient. They argued that the feedback might cause unnecessary concerns and overwhelm the clinicians with calls from worried patients. Furthermore, clinicians asserted that providing visual feedback of the cardiovascular data to the patients might not be meaningful due to the complex nature of the data, which patients might not understand. On the other side, all the patients were quite enthusiastic about being able to see their own ECG data and welcomed any feedback that the system can provide about their heart condition. P1 said that he would be interested in seeing what the data looks like. He would also like to get feedback from the app, but he would see the cardiologist for a diagnosis (if any). P2 argued that she would like to get some feedback, but that she would call the clinician if she suspected something to be wrong. P3, who had an anxiety disorder, argued that the feedback from the system could be helpful, especially for the patients with anxiety. She argued that; "it is better to know if your heart is normal than not knowing anything". Using the PCD approach, we found a compromise between these conflicting requirements and came up with a design requirement that the system should not provide any system-generated feedback that might worry the patient (such as automatic detection of AF), and at the same time, provide an overall visualization of selected data items.

3.3 System Design

Based on the requirement specifications, a *Minimum Viable Product* (MVP) focusing on three main features was outlined: (i) CVD data collection from a wearable ECG device, (ii) visualization of the data in the app, and (iii) collection of symptoms as reported by the patients. For the MVP, the Movisens 'EcgMove4' ECG device was chosen due to its availability, its data collection features, its

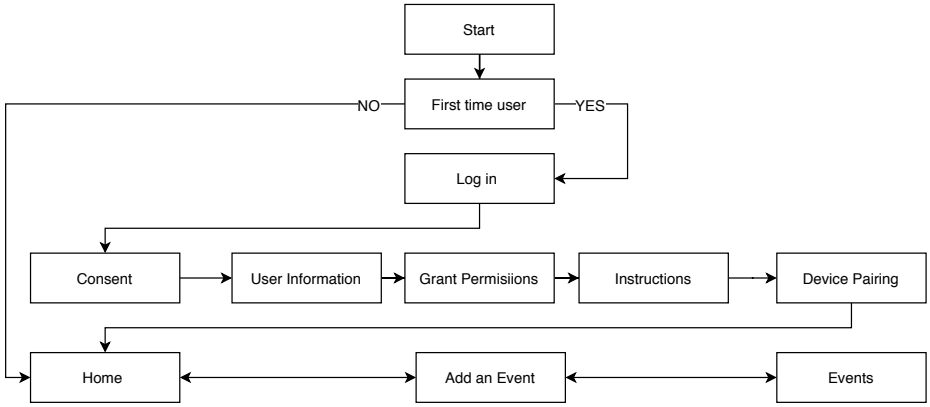


Fig. 4. The *mCardia* User Interface (UI) flowchart, including the user authentication and onboarding process, before the home screen (see Fig. 5) is shown.

connectivity, and its open application programmer interface (API). The EcgMove4 records ECG, HR, HRV, body position, step counts, and MET level, and allows patients to mark an event by double-tapping the device. The device can send data to a smartphone using Bluetooth low energy (BTLE). The system was designed and implemented during two major iterations.

3.3.1 Iteration 1. Initially, the overall user flow in the app was designed (Figure 4) with a set of mock-up designs for each screen. The first design iteration focused on the 'Home' and 'Events' screens in the app. Figure 5 shows the 'Home' screen. This is where all data is visualized and where the patient can get an overview. The overall design rationale of this screen is to provide an overview of relevant CVD physiological parameters, associated context information, and self-reported 'events' on a daily basis. This 24 hours overview is designed to be simple, aesthetically pleasing and informative. Therefore, it only shows selected data items that are most relevant for the patients to follow, namely; HR, HRV, sleep, active time, step counts, MET levels, and events. Other contextual and physiological data that is collected (e.g., raw ECG and accelerometer data, food-intake, body position, etc.) is not visualized.

An event can be reported in the system by pressing the plus (+) button or by double-tapping the ECG device. Event details are reported via the 'Event Details' screen (Figure 6a), which allow the patient to self-report symptoms, symptom duration, activity while the symptoms occurred, and a free-text note. The 'Events' screen (Figure 6b) lists the events reported by the patient, including both those entered using the plus button and the ones originating from a double-tap on the device. Events reported by a double-tap on the device is listed as '*Missing symptoms*', and the user can select this, and fill in the details on the details screen.

The initial prototype of the app was evaluated by the clinicians. Semi-structured interviews were used to assess the key aspects regarding relevance of the data collection and visualization. Clinicians liked the fact that the app does not provide any system generated feedback and agreed that the data visualized in the app is informative and that patients might be able to understand them. The feedback from the clinicians helped improve the overall design of the system, including, for example, how event details should be specified by the patient and the data visualization could be improved. The evaluation also revealed that some data were missing. For example, the daily intake of meals and beverages can impact hearth rhythms. All comments and feedback were analyzed and consolidated during the next design iteration.

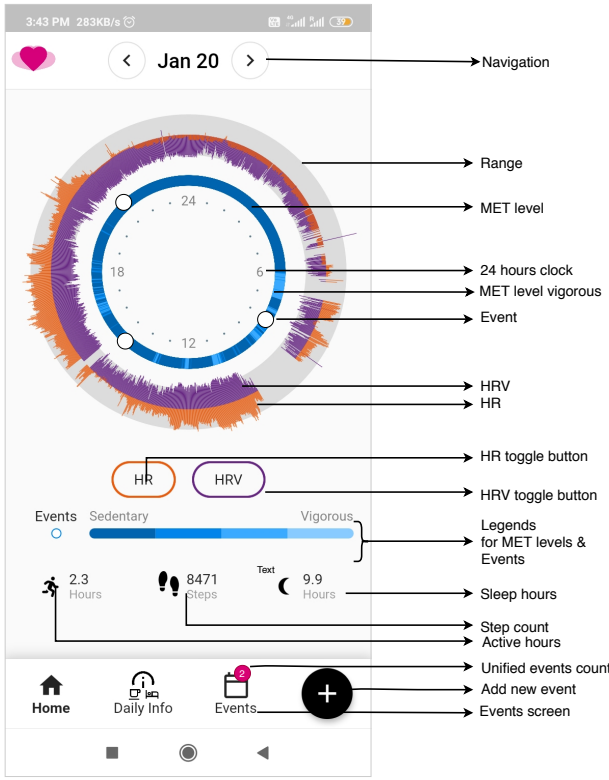
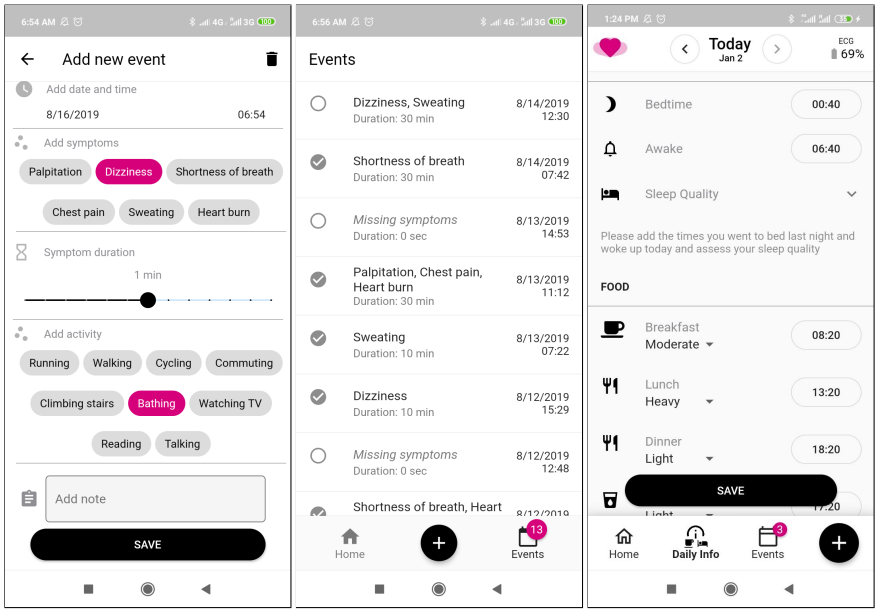


Fig. 5. The UI design of the *mCardia* home screen (final design). The main UI element is the 'wheel' which shows the detailed recordings of HR, HRV, and MET level in a 24-hour clock.

3.3.2 Iteration 2. Based on the feedback from the first iteration, an additional screen named 'Daily Info' (Figure 6c) was added. This page collects self-reported data on stress level, sleep quality, and dietary details like the timing and size of breakfast, lunch, and dinner. Finally, a set of introductory screens were added, which provide information to the patient on the first time the app is installed and used (see Figure 4). These screens include; (i) a set of informed consent screens, where the patient is informed about the purpose of the app and the study and can sign the consent form; (ii) a screen for collection of demographic data; (iii) a screen where the user grants permission to collect data from the phone (e.g., location data); (iv) a screen providing instruction for use; and (v) a screen instructing the patient on how to mount the ECG device on their chest and pair it with the phone.

Then the system was evaluated by three patients (P4: M/55; P5: F/60; P6: M/70) who used it for 24–72 hours on their own. Each patient was interviewed after using the system. The overall design and core features were well received. Specifically, patients stated that the visualization of the data on the app was helpful to see if the system was working. Patients argued that the system and continuous visualization increased their awareness of their health. For example, P5 said that the 'spikes' (HR and HRV) on the app was more interesting and informative than the heart rate data displayed in the Fitbit tracker they used. When asked how the visualization was informative to them, they said that the *mCardia* visualization provided them with an easy-to-understand, 24-hour visualization of when their HR/HRV was in or out of the range. The patients also provided input



(a) Event details (b) List of events (c) Daily Info

Fig. 6. The UI design of the *mCardia* mobile app (final design).

for improvements of the UI design of the app. For example, they wanted to be able to navigate back in time and see data from previous days, and they had suggestions for improving the legends. All these issues were incorporated into the final UI design as shown in Figure 5 and 6.

4 MCARDIA SYSTEM IMPLEMENTATION

Fig. 7 shows the overall software architecture of the *mCardia* mobile app, which runs on the patient's smartphone and collects passive and active contextual data. The *mCardia* app consists of a set of dedicated UI screens (marked red in Fig. 7), which implement the flow in Fig. 4 and the UI design in Fig. 5 and 6. mCardia is build using two frameworks; the CARP Mobile Sensing (CAMS) framework [6, 7] and the Research Package framework [27], which in turn consist of a number of sub-components (all marked in green in Fig. 7). CAMS is a cross platform and extensible framework for implementing mobile sensing apps and comes with a long list of options for data collection, data management, data anonymization, power (battery) optimization, and data upload. All data collection and data management in *mCardia* are handled by CAMS. The Research Package handles the informed consent flow, displays information about the study to the user, and asks for a signature.

The data sampling is configured as a 'Study' script in CAMS & handed over to the `StudyController`, which is then responsible for collecting and transforming the data according to the study specification. In *mCardia*, the data is stored in the CACHET Research Platform (CARP), a cloud-based infrastructure for managing and analysing mHealth data. The `CARPDataManager` is responsible for uploading data to CARP.

A set of sampling packages are registered with CAMS's study script, which are responsible for handling the data sampling. For example, all the contextual data collection (location, activity, and weather – see Table 1) is done via the `ContextSampling Package`. Similarly, step counts from the pedometer sensor in the smartphone are collected via the `SensorSampling Package`. Each

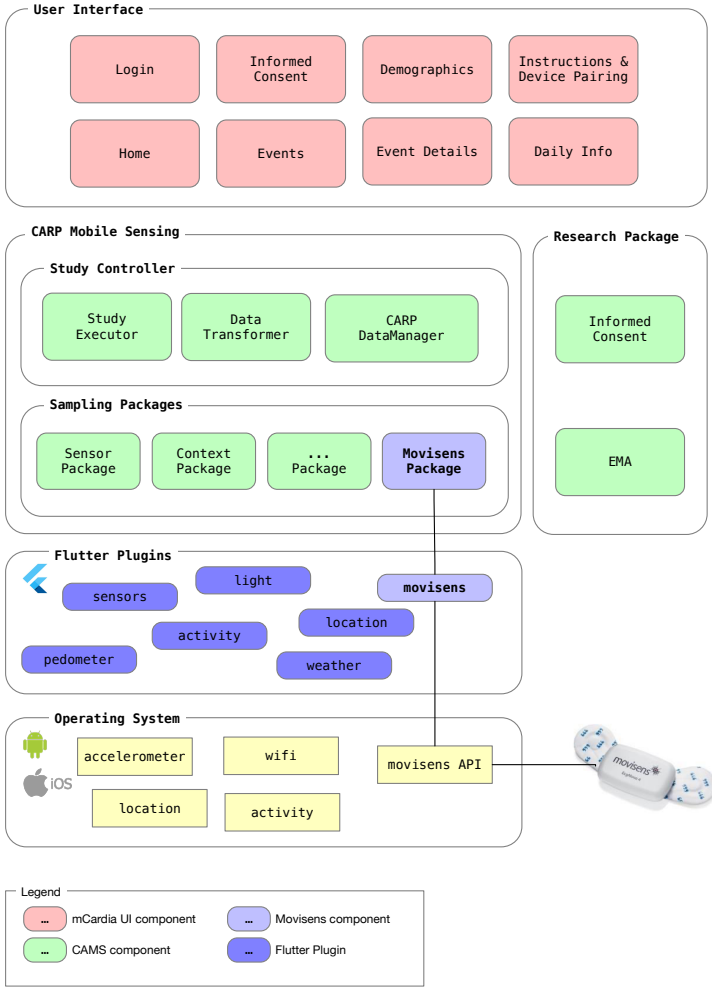


Fig. 7. Architecture diagram of *mCardia* app (red components) and its use of the CARP Mobile Sensing (CAMS) framework (green components) and Flutter plugins (blue components).

sampling package encapsulated access to the operating system (OS) sensors and typically uses one or more Flutter plugins to access the OS level sensors.

Likewise, the integration of the ECG Holter (Movisens EcgMove4) in *mCardia* was performed by implementation of a `MovisensSampling Package` [34], which usage a `Movisens` Flutter plugin [35] to access the native Movisens API via BTLE (all marked purple in Fig. 7). However, due to the extensible plugin architecture of CAMS, any new ECG Holters can be used as a plug-and-play by simply registering its sampling package with CAMS, without changing anything in the app itself.

4.1 ECG Sensor Data Management and Synchronization

As shown in Fig. 7, the ECG Holter's (Movisens in this prototype) Flutter plugin connects to it's native Android/iOS API over BTLE. It is capable of recording continuous ambulatory ECG with adhesive electrodes or a dry electrode textile chest belt, hence avoiding the need for cables. In

combination with the ECG sensor, the Holter also has on-board 3D accelerometer, gyroscope, barometric air pressure, and temperature sensors. Table 1 lists all the data types provided by the sensors, along with their sampling frequencies.

The Movisens Holter supports the 'live' and 'live + buffed' modes to communicate with these on-board sensors. In 'live' mode, the sensor signal can be activated via GATT notification, and if a sensor is not connected to a receiving device, then data is discarded. Therefore, this mode is not suitable for longitudinal data collection, where device disconnection is a very common scenario. For this reason, *mCardia* utilizes the 'live + buffer' mode in which sensors' signals are activated via GATT indication. As long as *mCardia* app remains connected, the device transmits data, and when disconnected, the data is buffered. It has a maximum buffering capacity of one day. On re-connection, the device sends the buffered data sequentially until all data is transmitted (or the connection is terminated).

The Movisens Holter records a one-lead 12-bit resolution ECG at a sampling frequency of 1024 Hz. The other sensors, like the 3D accelerometer and gyroscope sensors, are sampled at 64 Hz. In 'live' mode, the delay between measurement and the transmission of heart rate, heart rate variability, step counts, body position, and metabolic levels over BTLE is 70, 94, 94, 94, and 94 seconds respectively. Due to the high resolution and sampling frequency, the raw ECG recordings are not transmitted via BTLE for power saving reason. By default, all the data is also stored in Movisens Holter's 4GB internal memory. It has a capacity of holding of 14 days of raw ECG data. This ECG data is extracted manually by connecting the device to a PC via USB and then uploaded to the CARP cloud server, where the ECG and the other contextual data (collected via smartphone) are synchronized.

4.2 Data Visualization and User Information

mCardia visualizes data in real time as per the sampling frequencies in Table 1. Data visualization UI elements listen to the CAMS `StudyController` event stream, which broadcasts all data collected in real time. Data is not stored locally but uploaded directly to CARP. If the user wants to navigate back in time, data for a day is fetched from CARP and visualized. Sedentary behavior and active times are calculated based on the MET level data from the Holter device. Sleep is calculated once per day based on body position, heart rate, and other sensor data from the phone.

When the app is installed and used for the first time, it provides GDPR-compliant information to the user and shows a consent form that can be signed on the smartphone display, which is then stored in CARP. Upon installation, it also asks for the patients' demographic information, including gender, age, height, weight, and sensor location. The Movisens Holter uses this information for personalizing the computation of MET level, energy expenditure, etc. Since *mCardia* is designed to be used in an ambulatory setting (e.g., at home) for longitudinal data collection, the patient needs to understand how to set up and take care of the system. For this purpose, *mCardia* includes an elaborate tutorial that explains; (i) how to place the sensor and electrodes properly, (ii) how to clean the skin before placing the electrodes, and (iii) how to register an event by double-taping the sensor and annotate event in the app.

5 FEASIBILITY STUDY OF MCARDIA

Klasnja et al. [33] recommended that in the early phase of design or evaluation of novel health technologies *"a deep understanding of the how and why of the system use by its target users should be a central goal for evaluations of systems"*. Therefore, adhering to the best practices in health technology design research, a single-arm feasibility study of *mCardia* was carried out to obtain a comprehensive understanding of its usability and feasibility under free-living conditions. We applied the CACHET Unified Methodology for Assessment of Clinical Feasibility (CUMACF) [4, 8]

Parameters	Type	Source	Sampling rate
ECG	S	EcgMove4	1024 Hz
HR	S	EcgMove4	1/60Hz
HRV	S	EcgMove4	1/60Hz
MET Level	S	EcgMove4	1/60Hz
Acceleration	S	EcgMove4	64Hz
Rotation rate	S	EcgMove4	64Hz
Body position	S	EcgMove4	1/60Hz
Activity	S	Phone	EB
Steps	S	EcgMove4 & Phone	1/60Hz & EB
Events	PR	EcgMove4 & Phone	EB
Weather	S	Phone	4/day
Location	S	Phone	EB
Sleep	PR & S	Phone	1/Day
Noise level	S	Phone	1/120 Hz
Dietary	PR	Phone	1/Day

Table 1. Data features collected in *mCardia* with source and sampling rate. S: Sensed. PR: Patient-reported. EB: Event-based. Dietary includes food timings and type (light, moderate or heavy), sleep quality, and self perceived stress levels

method, which is an adoption from the Post Study System Usability Questionnaire (PSSUQ) [37] scale, Behavior Change Wheel Methodology [47], and Unified Theory of Acceptance and Use of Technology (UTAUT) [58] for assessing the user's intention for future acceptance of the technology. For a more comprehensive understanding, this was followed by the post-study semi-structured interviews. All the CUMACF [4] questions used in this study are available in the supplement. Specifically, the following three aspects of *mCardia* system were investigated:

- Usability evaluation which includes perceived user engagement, usefulness, and usability of *mCardia* in longitudinal ECG data collection.
- Technical robustness and feasibility of *mCardia*.
- Clinical usefulness of contextual data in the arrhythmia screening process.

5.1 Recruitment

Participants were recruited from Denmark and India. In Denmark, ethical approval for the study was obtained from the Danish research ethics committee (Journal-nr.: H-19071015). In India, approval was obtained from the institutional ethical committee (IRB) of our collaborator institute Mahatma Gandhi University of Medical Sciences & Technology (MGUMST), Jaipur.

In India, participants were recruited during their outpatient clinic visit at MGUMST's heart arrhythmia clinic. Likewise, in Denmark, participants were recruited by invitation and announcement to participate in the study outside the outpatient clinic, making it clear that observations in the study would be used for technical development, not for their clinical assessment or treatment. The study participant received a copy of the Q&A document describing the purpose of the study and the working of *mCardia* system. Inclusion criteria were; (i) previously undiagnosed individuals interested in heart arrhythmia screening; (ii) individuals who are already diagnosed with AF but are interested in tracking AF symptoms; (iii) comfortable in using smartphone apps and wearables or have a caretaker or family member who can help them in using *mCardia*; (iv) willing to use

mCardia for a minimum of two weeks. We used a rolling recruitment strategy for four months. A total of 33 participants were interested and met the inclusion criteria. All the participants signed the informed consent in the *mCardia* app.

5.2 Procedures

The study was divided into three phases; an orientation meeting, the ambulatory use of *mCardia*, and an end meeting. At the orientation meeting, a kit was provided to the study participant, containing a single-channel Movisens ECG device, a phone (if they needed), charging cables, and a user manual explaining how *mCardia* and the Movisens ECG device should be used. Participants were asked to sign a consent form in the *mCardia* app and provide some details regarding demography, medical history related to heart diseases, medication, and previous ECG monitoring experience. Besides, they were instructed on how to use the app and take care of the ECG device (charging, putting on, and taking off electrodes). During the study, participants were asked to continuously wear the ECG device for two weeks (or longer if they wanted to) and actively use the *mCardia* app for reporting and annotating any symptoms. They were also instructed to fill in the self-reports on food consumption, self-perceived stress, and sleep quality in the *mCardia* app on a daily basis. At the end of the study period, participants were asked to answer the CUMACF [4] questionnaire to evaluate the usefulness and usability of *mCardia*. Based on the participant's response to these questioners, we did a semi-structured qualitative interview for in-depth understanding of the experience in using *mCardia*.

No of participants	24
Sex – Female / Male	8 / 16
Age – Mean (SD)	58.79 (10.11)
Prior Holter experience	10
Assisted by caregivers	9

Table 2. Participants demographics.

6 RESULTS

Of 33 recruited participants, nine dropped out of the study and did not finish the minimum 2-week study period. The dropout reasons included; sudden deterioration of their health causing hospitalization ($N = 2$), skin allergy/irritability ($N = 4$) caused by the wet ECG electrode, and travel constraint ($N = 3$). Thus, we present the usability and technical feasibility of *mCardia* based on the data collected from 24 participants (8 females, 16 males, the average age of M = 58.79, SD = 10.11) over two weeks. Except for one participant (who used *mCardia* for over a month), all participants contributed equally (2 weeks) to the quantitative and qualitative data. All participants or their caretakers in this study reported owning a smartphone. However, we provided an Android phone to four participants as the *mCardia* was not compatible with their phone's older Android version. Total ten ($N = 10$), participants had previous experience of short-term (1-2 days) ECG screening using a traditional wired Holter monitor at home or in a hospital. In addition, 9 out of 24 participants were assisted by a family member or caretaker at home during the two week study period.

6.1 User Experience

6.1.1 Perceived Usefulness and Usability. Figure 8 shows the results of the CUMACF questions for perceived usefulness and usability. Overall 96% of the study participants responded very positively

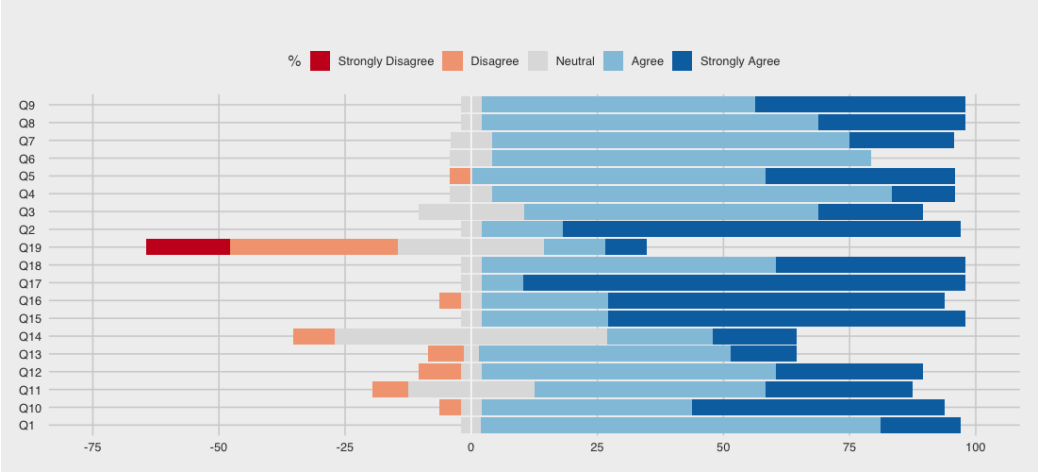


Fig. 8. Perceived usefulness and usability of *mCardia*. The list of CUMACF [4] questionnaire (Q1–Q19) is provided in the appendix.

about the design and usability of *mCardia* (Q1). They found the app to be unobtrusive and non-interfering in their day-to-day tasks. Nearly 95% of the participants reported that keeping track of daily activity and unusual symptoms can help them understand their symptoms and health better (Q5). Besides 78% found that *mCardia* could help increase the quality of communication with the doctor (Q3) and reduce recall bias (Q4) in home-based monitoring. Especially if they were to use the system for longitudinal arrhythmia screening. One participant who had previously used traditional wired Holder monitoring remarked that:

> *"The big wired Holter monitor was just a black box for me, and it was not very comfortable to sleep or work while wearing it. Also, I was not strict about keeping the symptoms diary, as I would not keep the diary and a pen with me at all time."* [P23]

In the "Effort Expectancy" assessment, nearly 75% of the participants found the help, error messages, guidelines etc. offered in *mCardia* adequate (Q11). In response to the question whether *mCardia* has all the functionality expected (Q14), around 50% of the participants responded neutrally, whereas, 9% showed disagreement. During the post-study interview, we found that participants wanted support for medication tracking and reminders in *mCardia*, since keeping track of their medication was core to their treatment. Finally, when asking about "Facilitating Conditions" most participants were positive. They reported that they had the resources needed (i.e smartphone) (Q17, N=95.8%) and assessed themselves to be skillful in the use of *mCardia* (Q18, N=95%). Besides, only 8.3% reported that they would require a dedicated person to assist (Q19).

Overall, the *mCardia* was reported as easy to use, with high user satisfaction, especially amongst the participants who have had previous experience with traditional wired Holter monitoring; 8 of 10 these users were very satisfied.

6.1.2 Engagement as time spent. The amount of time spent on the mobile app is another indicator of user engagement. On average, participants spent 21 minutes interacting with *mCardia* daily. We found the use pattern for filling the details about recorded symptoms or events to be random. However, daily information such as stress level, sleep quality, and food intakes were mostly filled in once in the evening. Besides these two tasks, the majority of the participants routinely interacted with *mCardia* in the morning (to check night trends in heart rate) or after performing physical

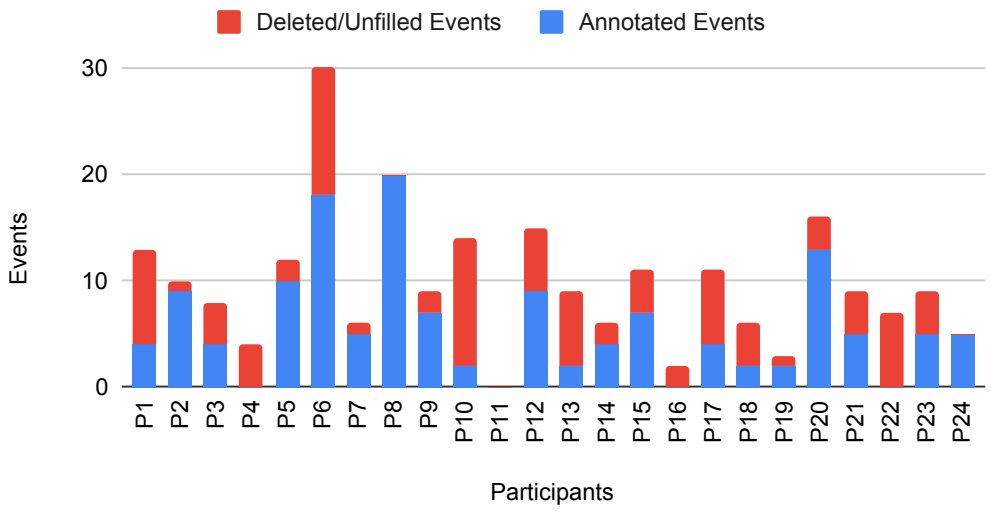


Fig. 9. Number of annotation events and number of deleted or unfilled events per participant.

activities such as exercise, cycling, or a long walk. It should be noted that *mCardia* was not designed to maximize the time spent on the app. As one participant states in the post-study interview:

> *"I might have spent more time in the app, if it had provided additional features such as medication tracking or recommendations."* [P7]

However, based on the initial design phase, *mCardia* was designed mainly for brief sporadic use and not for medication management.

6.1.3 Events recording and phone-based context collection. A total of $N = 235$ events were created (either by taping on the Movisens device or on the phone), out of which nearly 60% were annotated, and the other 40% were either deleted or remained unfilled. Figure 9 shows the number of annotated and deleted/unfilled events per participant during the study. Among all the participants, the frequency of events created and deleted due to accidental taps on the ECG device was more on the initial few days and declined as the participants became more familiar with *mCardia*.

6.2 Performance of data collection and its impact on usability

During the feasibility study over 8064 hours of contextualized ECG data from 24 participants was collected. Nearly 89% of the collected ECG data is suitable for arrhythmia analysis. The remaining 11% data is not usable primarily because the patient forgot to put on the device, or because the electrodes became nonadhesive resulting in poor signal quality. We also found that the percentage of unusable data was slightly higher among the nine participants who were assisted by family members or caretakers. During the post-study interview, we learned that when electrodes became nonadhesive, they would not realize this and replace them, unless told by the caretakers. The caretakers could spot missing HR and HRV data in the *mCardia* app. One caretaker recalled:

> *"I would usually check and change the electrodes only when I found gaps in the HR or HRV data in the mCardia app's circular wheel."* [P3]

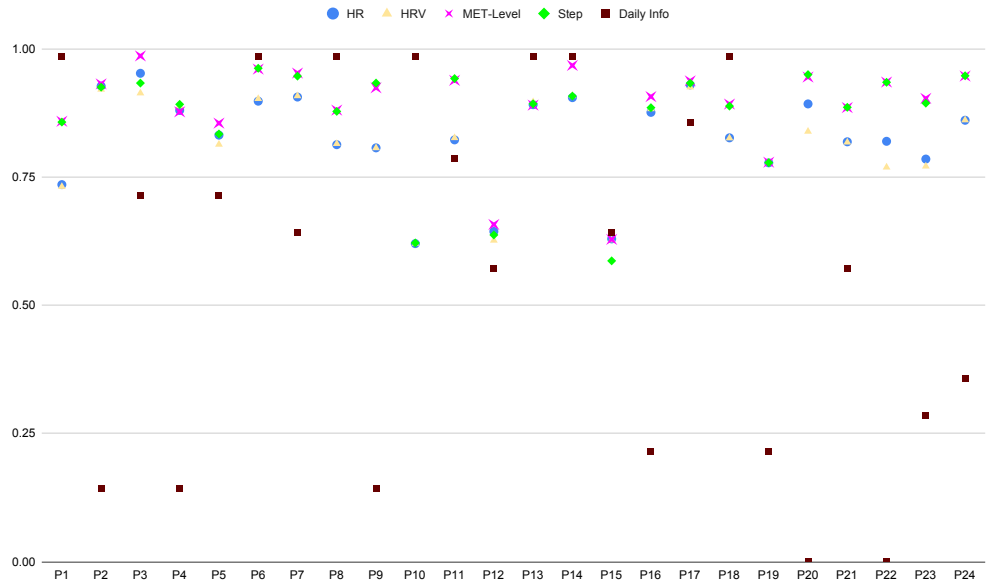


Fig. 10. Data collection yield per participant

6.2.1 Yield. Li et al. [38] define yield as "the fraction of the expected samples to the actual number of samples collected by the system" and argue that yield can serve as a proxy for determining engagement and user compliance. Following this definition, we define yield in *mCardia* as the fraction of the expected samples of various data points such as HR, HRV, MET-level, and Daily Info that are successfully collected and stored in the cloud back-end. For instance, since HR, HRV, MET-level (corresponding to light, moderate, and vigorous activities), and steps are collected once per minute, 60 samples of them are expected in any given single hour. Similarly, Daily Info is collected once per day; therefore, the fraction of days in which we collected Daily Info data defines the Daily Info yield. Figure 10 shows the yield of HR, HRV, MET level, Step, and Daily Info for each participant. The median yield of Daily Info is 0.671, and only 57% of participants have Daily Info yield higher than 0.6. The MET-level median yield is 0.903, and nearly 88% of participants reaching yield more than 0.85. Similarly, the median yield for step is 0.902, and 81% of participants reaching yield more than .85. In contrast, the median yield of both HR and HRV is 0.79, which is lower than steps and MET-level, and only 41% and 37% of participants have HR and HRV yield higher than 0.85, respectively. The potential reasons for low HR and HRV yield include; (i) user not wearing the ECG device or wearing discharged device, (ii) non-adhesive of the ECG electrodes over time resulting in a noisy signal, and (iii) arrhythmia episodes in which the Movisens' on-board algorithm is unable to correctly calculate HR and HRV.

6.3 Qualitative Data

To further understand the participants' perspective, we conducted a post-study semi-structured interview of each participant and their caretaker. We employed an inductive approach [57] for analyzing this interview data. As the focus of this study was on the feasibility and perceived

Factors	
Total duration of ECG recording	8054 Hrs
Total events events registered	235
Events deleted or unfilled	98
Total annotated events	137
Average no of events filled per participants	5.7

Table 3. Overall collected ECG and contextual data

usefulness of *mCardia*, the following themes were in focus; (i) issues encountered, (ii) task suitability, and (iii) perceived usefulness.

6.3.1 Issues encountered by participants. Table 4 lists the four main issues reported by the study participants during the post-study interview. One-fourth of the participants reported issues related to non-adhesiveness of wet ECG electrodes due to sweating or body movements during the night. This mainly happened when participants forgot to change the electrodes daily and continued using the same electrodes for more than 24 hours. A related issue is the skin discomfort or irritability which was reported by 20% of the participants – an issue that is well-known in long-term ECG monitoring [45]. A third issue reported by 38% of the participants was the creation of "false events" due to accidental tapping on the ECG sensor. Such accidental taps typically happened when putting the device on or off for charging. Such taps resulted in creating false events in the *mCardia* app and it annoyed the participants as they had to delete them manually in the app. One participant recalled:

> *"For the first two days, when I pushed the ECG device hard in order to fit it into the charging tray, it added some events. When I saw these empty events later in the app, I was confused as I didn't recall tapping on the device."* [P10]

Finally, 20% of the participants reported that they kept wearing the ECG sensor even when its battery was completely discharged. Although the *mCardia* app displays the battery level of the sensor, the participants said that they did not notice it unless they explicitly opened the app and looked at the battery level. As explained by a participant:

> *"If the app is closed and the device battery dies, I would not realize it. I [would] continue wearing it for hours until I opened the app and looked at the battery level. This happened several times in the last two weeks."* [P17]

6.3.2 Task suitability. From the user's point of view, the main task was to collect high-quality event information during ambulatory long-term use. Hence, we investigated the feasibility of phone-based event annotation as compared to the traditional paper-based diary. Compared to traditional paper-based event diaries, participants found the event creation and annotation in *mCardia* simple, clean, and very convenient. Interestingly, phone-based event annotation was most liked by participants ($N = 10$) who previously have had the experience of traditional home-based Holter monitoring and paper-based event diaries. As explained by P20:

> *"It is easier to note the symptoms on the mobile phone, since I carry it all the time. I do not have to remember and write it down in my [paper-based] event diary – especially, if I have to do it for many days or months."* [P20]

6.3.3 Perceived Usefulness. Although context-aware ECG monitoring via *mCardia* is primarily designed to help cardiologists in a more accurate and informed assessment of arrhythmias, we

Issue	%
Electrodes becoming non-adhesive	26%
Skin irritability due to continuous use of wet ECG electrodes	20%
Accidental tapping on ECG device causing false event registration	38%
ECG device battery discharged	20%

Table 4. Issues identified and their relative distribution.

were also interested in the participants' perceived usefulness of *mCardia* for managing their CVD health. In this regard, several participants or their caretakers described that they believed *mCardia* could be useful in two ways; (i) in better communicating their symptoms to the cardiologist, and (ii) to keep them aware about any unusual symptoms and the context in which they appear. For example, as explained by P2:

> *"I think this clock overview is nice. I can see how my heartbeat changes when I am doing different activities. On two Fridays – when I had been playing basketball – I felt palpitations during the night and registered these events. It makes it easy to show them to the cardiologist and ask what happened at that time, and if it is related to playing sports."* [P2]

6.4 Clinical case studies

The usefulness of context-aware ECG monitoring in arrhythmia screening can be demonstrated in two clinical cases based on data collected from the feasibility study – one focusing on tachycardia (high restring heart rate) and the other looking at palpitations.

6.4.1 Clinical case #1: Tachycardia. Tachycardia is the condition in which the heart rate exceeds the normal resting rate, which can be physiological or due to abnormal heart rhythm, then called tachyarrhythmia. The definition of tachycardia in adults is a resting HR above 100 BPM [3]. Figure 11 (a) and (b) shows a sudden increase in heart rate to over 100 BPM on two different occasions for the same patient (P12, female, 70 years). In Figure 11(b), contextual data reveals that the subject is not doing any physical activities during this time or even prior to the onset. During this period, the participant also added an event by tapping the device and provided some additional context that she was lying down after dinner and felt severe heartburn symptoms that lasted for around 20 minutes. This makes it a typical case of a supraventricular tachycardia (SVT) episode. On the other hand, Figure 11 (a) although showing HR above 100 (at around 10-10:30 PM), contextual data shows that the patient is doing physical activity of jogging, walking, and running prior and during this period. Hence, in this scenario HR above 100 is not a case of SVT.

This case demonstrates that interpretation of ECG segments and HR is different depending on the context, and assessment in isolation could potentially lead to misdiagnosis. Thus, collection of contextual data in ambulatory ECG monitoring could enable faster and more accurate assessment of arrhythmias in both manual and computer computer-aided diagnosis of arrhythmia.

6.4.2 Clinical case #2: Palpitations. In this case, the patient was evaluated for annoying palpitations, and had known permanent atrial fibrillation. He referred palpitations during sleep that lasted 30 minutes. Figure 12 (a) shows cases of mild rate changes whereas the accelerometer shows that patient is just turning in bed. At this point his doctor may hesitate to increase rate-lowering medications since the patient also referred dizziness when changing from laying down to standing position. In this context, Figure 12 (b) helps to choose an adequate medicine adjustment since it is

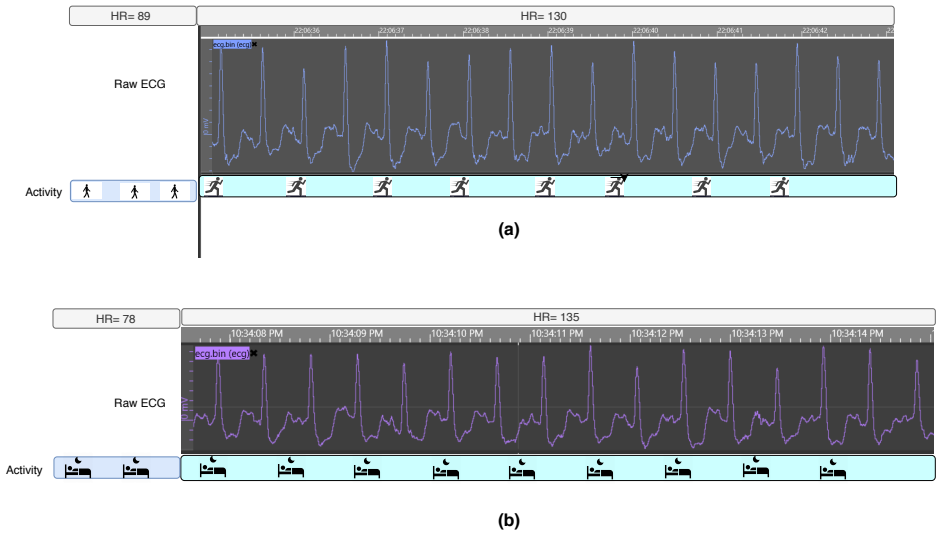


Fig. 11. Clinical case #1 demonstrating the usefulness of context information in ECG interpretation. Although both ECG snippets from the same participant have HR >130, the context information (sleeping or running) helps in distinguishing that (b) is a case of supraventricular tachycardia (SVT), whereas (a) is normal.

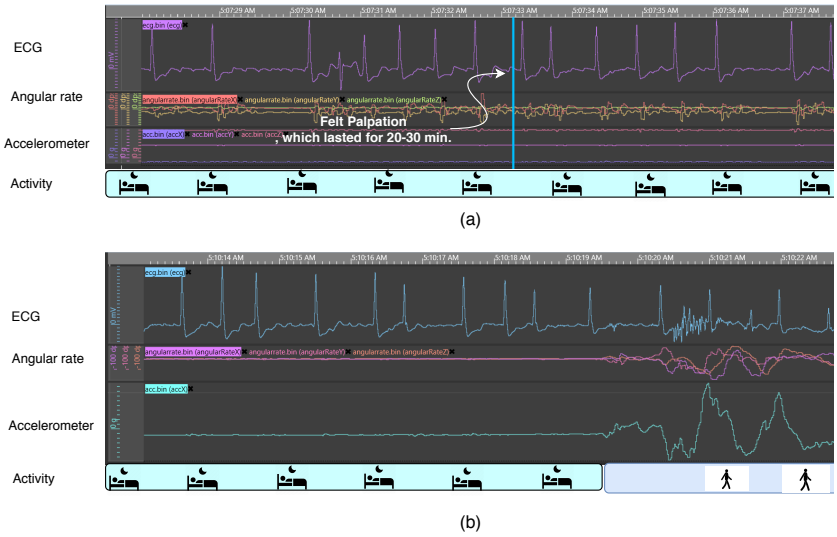


Fig. 12. Clinical case #2 demonstrating the usefulness of context information in medication prescription. (a) Subject tapped on device and registered an event. (b) Subject gets up from the bed and starts walking.

reassuring that there is an adequate chronotropic response when changing from laying in bed to walk, suggesting problem with ortostatic blood pressure changes. Furthermore, there are no case of

severe low rate. Hence, choice could be made for a rate-lowering medicine without blood-pressure lowering effect. This demonstrates that, in this case collection of contextual data not only helps in diagnosis but also in deciding on correct medication treatment.

7 DISCUSSION

This study aimed to test the feasibility and usability of *mCardia* for collecting context-aware ambulatory ECG for arrhythmia screening under free-living conditions. To this end, we need to ensure that *mCardia* meets the user experience and usability standards and keeps participants engaging as well as informed. The overall results of this feasibility study are encouraging. The findings indicated a good degree of usability, usefulness, and clinical applicability of *mCardia*. With the aim of replacing wired Holter monitors and paper-based diaries, it is especially interesting to notice that the quantitative results on usability and feasibility were more positive from the participants ($N = 10$) who had previously used a traditional setup for home-based ECG monitoring. In this section, we discuss the patient's perspective on the *mCardia* system, areas of improving it, and the use of contextual information in arrhythmia analysis.

7.1 Patient's perspective on *mCardia*

From the users' perspective, three interesting findings emerged from the study, namely that *mCardia* could improve event reporting practice, could improve patient clinician communication, but that medication tracking was missing.

7.1.1 Improved event reporting practice. The ability to fill the details of the symptomatic events on the phone rather than a paper-based event diary was especially liked by the participants who previously had undergone 1-2 days ambulatory Holter monitoring at home. One participant noted that;

> *"It was much easier to remember that I had tapped the device and had unusual symptoms by looking at the unfilled event log in the mCardia app. In my previous home Holter test, I rarely maintained the event diary, and even when I did the entries, it was with an approximate time."*[P5]

Careful mapping of an event's timestamp to the ECG timestamp is vital because – as shown in clinical case #2 – only accurate mapping of these events helps in a better understanding of these unusual symptoms during ECG analysis. If relied on recall memory and proximate timestamp, these reported symptoms might not be reflected in corresponding ECG, which may cause ambiguity during analysis. By tapping the ECG device and using the phone to provide the details of the event significantly reduced recall bias in reporting events.

7.1.2 Improvement in patient clinician communication. As reported in the qualitative findings (Section 6.3), the patients argued that *mCardia* has a strong potential to improve communication about their health condition to the doctor in a longitudinal home-based arrhythmia screening. In particular, the visualization of the daily overview of HR with the different activities and symptomatic events was considered motivating enough to keep using *mCardia* for a longer period. Although participants asked if *mCardia* could give them real-time feedback on arrhythmia, this feature was intentionally not part of the design. During the PCD design process, the doctors warned against such a feature. Current machine learning algorithms for automatic arrhythmia detection are still in their infancy and mainly work on high-quality ECG recording done in the clinic under controlled conditions. Therefore, enabling real-time automatic arrhythmia detection on ambulatory ECG could result in false positives, causing anxiety and unnecessary hospital visits. This problem has

also been reported in several other studies involving short (30 sec) ECG even in a low-risk healthy population [30, 50].

Instead, *mCardia* can be seen as a tool for improving patient-clinician communication during the longitudinal arrhythmia screening period. On the one hand, it helps collect a much richer data set for the clinician to improve clinical decision-making. On the other hand, it provides the patient with much better awareness and understanding of their disease symptoms and the relationship between behavior and their heart-related symptoms, without providing the diagnosis. Taken together, this can improve the dialogue between the patient and clinician.

7.1.3 Medication tracking and reminders. In CUMACF questionnaire, question no. 14 ("Q14: mCardia has all the functionalities that I expect it to have?") received the most negative response, with a majority of participants disagreeing to this statement (see Figure 8). In the post-study interview, we learned that the majority of participants who disagreed to this statement wanted medication tracking and reminders to be part of *mCardia*. This was especially highlighted by participants who suffered from several co-morbid chronic conditions. Even though medication tracking was discussed during the *mCardia* design process, we decided not to include this since there are plenty of other apps available for this. However, the study showed that this might be important to *mCardia* after all. Partly because the study showed that participants wanted this as an integrated feature in the same app. But more importantly, because medication tracking could help understand if there are any correlations between the medication dose and timing, and the occurrence of arrhythmia events or unusual symptoms experienced by participants. Furthermore, medication tracking could help in keeping participants more engaged in the data collection process.

7.2 Improvements to the *mCardia* system

While *mCardia* was widely accepted, we found inspiration for several areas of technical improvements concerning improving signal quality, keeping the ECG device charged, the prevention accidental event logging, and improvement of self-reported data.

7.2.1 Improving signal quality. Proper attachment of the electrodes to the skin as well as changing them daily is key to quality transmission of the ECG signal. Support for continuous monitoring of signal quality could be added to *mCardia* which could provide in-app notification if the signal quality remains weak for more than a certain period. Moreover, motion sensors, such as accelerometers, makes it possible to distinguish whether poor signal quality is due to motion artifacts or due to non-adhesiveness of the ECG electrode.

In this context it is worth noticing that compared to others, the participants assisted by caretakers (N=9) have a high percentage of unusable ECG data and low HR and HRV yield. The high percentage of unusable data is because caretakers would leave the phone with the patient and mainly check *mCardia* in the morning and evening. If electrodes become nonadhesive, caretakers would not realize this and replace them until they open the mCardia app on participant's phones and see the missing data. To overcome this challenge, *mCardia* could also send a notification on signal quality to the caretaker's phone.

Electrodes are also one of the common causes of skin discomfort and a major reason for participant dropout in ECG monitoring studies [45]. Our study was not exempt from this. Although the participants were instructed to change the electrodes every day, some patients did not adhere to this instruction resulting in skin irritability and eventually dropping out of the study. One way to address this challenge could be to implement a daily notification in the app and alert the patient to change the electrodes. This feature could be combined with signal quality reporting.

7.2.2 Keeping the ECG device charged. We found that although the battery level of the ECG device was displayed in the *mCardia* app, participants often forgot to charge the device. This could result in full discharge of the device and participants could end up wearing a discharged device for several hours. They would notice this only when they open the app. This was especially prevalent in the ($N = 9$) participants who were assisted by a caretaker. This problem could be mitigated if *mCardia* notified the user or caretaker with an alert on low battery level, even if the app is closed. This can be done by server-sided push notification based on the last charging status of the ECG device and its discharge rate, which would work even if the *mCardia* app is closed or if the device becomes disconnected.

7.2.3 Preventing accidental event logging. Unintended logging of events due to accidental taps on the ECG device, especially while putting on or taking off the device to charge, was one of the main issues surfaced in qualitative analysis. Although detecting "accidental events" could be device-specific (Movisens in this case), they can be prevented by checking whether ECG electrodes are attached to the device or not. When the electrodes are not attached to the device, and tap on it should simply be discarded. In this way, new events will only be listed in the *mCardia* app when a participant is wearing the device and the recording is in progress.

7.2.4 Improving reporting of daily info. The information such as stress, sleep, and food intake provides important contextual and behavioral information that supplements the underline ECG data during arrhythmia analysis. However, the relatively low yield of Daily Info (c.f., Figure 10) suggests that participants were not motivated to enter these details on a daily basis. When asking about the reason for this low yield, we learned that participants could not see any direct value of this information for themselves. The design decision to collect this information was primarily focused on its importance for clinicians during ECG analysis. Hence, it seems important to feed this information back to the user as day-to-day educational information about the potential linkage between lifestyle and commonly known arrhythmia triggers.

7.3 Using contextual information in arrhythmia analysis

The two clinical cases highlighted in this study demonstrate the usefulness of contextual data in making more clinically informed decisions when analysing ambulatory ECG for arrhythmia diagnosis. For these case studies, the temporal alignment of ECG with the different contextual data was done manually. To take full advantage of these different contextual data, a fully automated system for temporal alignment, synchronization, and visualization of the context and ECG would be needed. Moreover, semi-automatic detection of potential arrhythmia onsets and offsets time would be needed to quickly analyze such a large amount of data.

In addition, contextual features can be used together with the ECG morphology to achieve personalization in algorithms for arrhythmia detection. Arrhythmia-provoking user-contexts and common symptoms at the onset of arrhythmias (as listed by Hansson et al. [24]) can be utilized to build context-specific heuristics, which can help in reducing false alarms in computer-aided diagnosis of ambulatory ECG.

8 LIMITATIONS AND FUTURE WORK

Our sample size is small, which might limit the ability to generalize the findings. Also, as there is no optimal length of the arrhythmia screening period and the study was focused more on technology demonstration or proof-of-concept rather than clinical outcomes, we kept the study period to a minimum of two weeks. However, it would be interesting to learn how the participant's response, engagement, and usability behavior changes when they use mCardia for a much extended period (i.e., 2-4 months).

In the next iteration, we would like to address the issue highlighted in this study namely: (1) provide alerts to users or caretakers on signal quality when electrodes become non-adhesive, (2) add medication tracking and reminders to increase user engagement, and (3) provide alerts to users when the ECG device battery needs charging, even while the mCardia app is closed or running in the background. The plan is to use and deploy *mCardia* in a larger-scale clinical trial as part of the REAFEL project [5] for diagnosis and management of Atrial Fibrillation in frail and elderly patients.

9 CONCLUSION

This paper has described the design, technical implementation, and initial deployment of *mCardia*, a context-aware longitudinal ambulatory ECG collection system for cardiac arrhythmia screening. The primary contributions of the work are threefold. First, we have identified the relevant contextual information that can help improve arrhythmia screening when combined with ECG data. Second, we have presented the design and technical implementation of a device-agnostic plugin-based mHealth application for collecting ECG with a broad spectrum of contextual data. Third, a non-randomized, single-arm feasibility study in a Danish and Indian setting demonstrated the usability and user acceptance of such technology for continuous ambulatory ECG monitoring for arrhythmia diagnosis. The *mCardia* system achieved high data yield, as well as high levels of patient compliance and acceptance. Via two clinical case studies, we also demonstrated how contextual information could help improve arrhythmia screening. As such, this paper presents promising results in terms of the usability and feasibility of *mCardia* system for longitudinal arrhythmia diagnosis and monitoring under free living conditions.

ACKNOWLEDGMENTS

First and foremost we would like to thank all the participants in this study. Mr. Bipin Kumar Rathod, Dr. A K Singh, and Dr. Rajeev Sharma at Mahatma Gandhi University of Medical Sciences Technology, Jaipur and Dr. Kamal Sharma, in Ahmedabad helped with the recruitment in the India part of the study. Andrea Q. Lopategui and Barnabás Homola helped implement the *mCardia* app. Darius Rohani helped with the data analysis.

REFERENCES

[1] AliveCor. 2020. AliveCor Inc. Retrieved July 8, 2020 from https://www.alivecor.com/kardiamobile
[2] Daejoon Anh, Subramaniam Krishnan, and Frank Bogun. 2006. Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis. *Journal of electrocardiology* 39, 3 (2006), 343–345.
[3] Eric Awtry, Cathy Jeon, and Molly G Ware. 2006. *Blueprints cardiology*. Lippincott Williams & Wilkins.
[4] Jakob E. Bardram. 2019. CACHET Unified Method for Assessment of Clinical Feasibility. Retrieved May 2, 2020 from https://www.cachet.dk/research/cumacf
[5] Jakob E. Bardram. 2019. REAFEL: Reaching the Frail Elderly Patient for optimizing diagnosis of atrial fibrillation. Retrieved May 2, 2020 from https://www.cachet.dk/research/research_projects/reafel
[6] Jakob E. Bardram. 2020. The CARP Mobile Sensing Framework – A Cross-platform, Reactive, Programming Framework and Runtime Environment for Digital Phenotyping. arXiv:cs.HC/2006.11904
[7] Jakob E. Bardram. 2020. CARP Mobile Sensing Framework in Flutter. Retrieved June 23, 2020 from https://pub.dev/packages/carp_mobile_sensing
[8] Jakob E. Bardram. 2020. Evaluating Personalized Pervasive Health Technology—But How? *IEEE Pervasive Computing* 19, 3 (2020), 37–44.
[9] Noureddine Belgacem and Selma Boumerdassi. 2009. Mobile personal electrocardiogram monitoring system with patient location. In *Proceedings of the 1st ACM international workshop on Medical-grade wireless networks*. ACM, 69–72.
[10] Emelia J Benjamin, Philip A Wolf, Ralph B D'Agostino, Halit Silbershatz, William B Kannel, and Daniel Levy. 1998. Impact of atrial fibrillation on the risk of death: the Framingham Heart Study. *Circulation* 98, 10 (1998), 946–952.
[11] Frank Bogun, Daejoon Anh, Gautham Kalahasty, Erik Wissner, Chadi Bou Serhal, Rabih Bazzi, W Douglas Weaver, and Claudio Schuger. 2004. Misdiagnosis of atrial fibrillation and its clinical consequences. *The American journal of*

medicine 117, 9 (2004), 636–642.

[12] Pak-Hei Chan, Chun-Ka Wong, Yukkee C Poh, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Ming-Zher Poh, Daniel Wai-Sing Chu, and Chung-Wah Siu. 2016. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association* 5, 7 (2016), e003428.

[13] Sumeet S Chugh, Rasmus Havmoeller, Kumar Narayanan, David Singh, Michiel Rienstra, Emelia J Benjamin, Richard F Gillum, Young-Hoon Kim, John H McAnulty Jr, Zhi-Jie Zheng, et al. 2014. Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation* 129, 8 (2014), 837–847.

[14] Sai Manoj Pudukotai Dinakarrao, Axel Jantsch, and Muhammad Shafique. 2019. Computer-aided Arrhythmia Diagnosis with Bio-signal Processing: A Survey of Trends and Techniques. *ACM Computing Surveys (CSUR)* 52, 2 (2019), 23.

[15] Marcus Dörr, Vivien Nohturfft, Noé Brasier, Emil Bosshard, Aleksandar Djurdjevic, Stefan Gross, Christina J Raichle, Mattias Rhinisperger, Raphael Stöckli, and Jens Eckstein. 2019. The WATCH AF trial: SmartWATCHes for detection of atrial fibrillation. *JACC: Clinical Electrophysiology* 5, 2 (2019), 199–208.

[16] Abdur Forkan, Ibrahim Khalil, and Zahir Tari. 2013. Context-aware cardiac monitoring for early detection of heart diseases. In *Computing in Cardiology 2013*. IEEE, 277–280.

[17] Abdur Rahim Mohamad Forkan and Weichich Hu. 2016. A context-aware, predictive and protective approach for wellness monitoring of cardiac patients. In *2016 Computing in Cardiology Conference (CinC)*. IEEE, 369–372.

[18] Curt D Furberg, Bruce M Psaty, Teri A Manolio, Julius M Gardin, Vivienne E Smith, Pentti M Rautaharju, CHS Collaborative Research Group, et al. 1994. Prevalence of atrial fibrillation in elderly subjects (the Cardiovascular Health Study). *The American journal of cardiology* 74, 3 (1994), 236–241.

[19] Toshiyuki Furukawa, Roberto Maggi, Cristina Bertolone, Fabrizio Ammirati, Massimo Santini, Renato Ricci, Franco Giada, and Michele Brignole. 2011. Effectiveness of remote monitoring in the management of syncope and palpitations. *Europace* 13, 3 (2011), 431–437.

[20] Alessio Galli, Francesco Ambrosini, and Federico Lombardi. 2016. Holter monitoring and loop recorders: from research to clinical practice. *Arrhythmia & electrophysiology review* 5, 2 (2016), 136.

[21] Perry M Gee, Deborah A Greenwood, Debora A Paterniti, Deborah Ward, and Lisa M Soederberg Miller. 2015. The eHealth enhanced chronic care model: a theory derivation approach. *Journal of medical Internet research* 17, 4 (2015), e86.

[22] Godwin Denk Giebel and Christian Gissel. 2019. Accuracy of mHealth devices for atrial fibrillation screening: systematic review. *JMIR mHealth and uHealth* 7, 6 (2019), e13641.

[23] Jan Gulliksen, Bengt Göransson, Inger Boivie, Stefan Blomkvist, Jenny Persson, and Åsa Cajander. 2003. Key principles for user-centred systems design. *Behaviour and Information Technology* 22, 6 (2003), 397–409.

[24] Anders Hansson, Bjarne Madsen-Härdig, and S Bertil Olsson. 2004. Arrhythmia-provoking factors and symptoms at the onset of paroxysmal atrial fibrillation: a study based on interviews with 100 patients seeking hospital assistance. *BMC cardiovascular disorders* 4, 1 (2004), 13.

[25] Jeffrey S. Healey and Roopinder K. Sandhu. 2015. Are We Ready for Mass Screening to Detect Atrial Fibrillation? *Circulation* 131, 25 (2015), 2167–2168. https://doi.org/10.1161/CIRCULATIONAHA.115.017288

[26] Gerhard Hindricks, Evgueny Pokushalov, Lubos Urban, Milos Taborsky, Karl-Heinz Kuck, Dmitry Lebedev, Guido Rieger, and Helmut Purerfellner. 2010. Performance of a new leadless implantable cardiac monitor in detecting and quantifying atrial fibrillation results of the XPECT trial. *Circulation: Arrhythmia and Electrophysiology* 3, 2 (2010), 141–147.

[27] Barnabás Homola and Jakob E. Bardram. 2019. Research Package – A Flutter Framework for mHealth Research. Retrieved June 23, 2020 from https://www.researchpackage.org

[28] Apple Inc. 2020. Taking an ECG with the ECG app on Apple Watch Series 4 or later. Retrieved July 8, 2020 from https://support.apple.com/en-us/HT208955

[29] Inc iRhythm Technologies. 2020. Zio- The Flexible, Patient-Friendly, Accurate Cardiac Monitor. Retrieved Jan, 2020 from https://www.irhythmtech.com/professionals/why-zio

[30] Nino Isakadze and Seth S Martin. 2019. How useful is the smartwatch ECG? *Trends in cardiovascular medicine* (2019).

[31] Chakkalakkal Prabhakaran Karunadas and Cibu Mathew. 2020. Comparison of arrhythmia detection by conventional Holter and a novel ambulatory ECG system using patch and Android App, over 24 h period. *Indian Pacing and Electrophysiology Journal* 20, 2 (2020), 49–53.

[32] Paulus Kirchhof, Angelo Auricchio, Jeroen Bax, Harry Crijns, John Camm, Hans-Christoph Diener, Andreas Goette, Gerd Hindricks, Stefan Hohnloser, Lukas Kappenberger, et al. 2007. Outcome parameters for trials in atrial fibrillation: recommendations from a consensus conference organized by the German Atrial Fibrillation Competence NETwork and the European Heart Rhythm Association. *Europace* 9, 11 (2007), 1006–1023.

[33] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 3063–3072.

[34] Devender Kumar, Thomas Nilson, and Jakob E. Bardram. 2020. CARP Movisens Sampling Package. Retrieved June 23, 2020 from https://pub.dev/packages/carp_movisens_package

[35] Devender Kumar, Thomas Nilson, and Jakob E. Bardram. 2020. Movisens Flutter Plugin. Retrieved June 23, 2020 from https://pub.dev/packages/movisens_flutter

[36] Craig Larman. 2012. *Applying UML and patterns: an introduction to object oriented analysis and design and interative development*. Pearson Education India.

[37] James R Lewis. 1992. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 36. SAGE Publications Sage CA: Los Angeles, CA, 1259–1260.

[38] Dingwen Li, Jay Vaidya, Michael Wang, Ben Bush, Chenyang Lu, Marin Kollef, and Thomas Bailey. 2020. Feasibility Study of Monitoring Deterioration of Outpatients Using Multimodal Data Collected by Wearables. *ACM Transactions on Computing for Healthcare* 1, 1 (2020), 1–22.

[39] Jin Peng Li, Damon Berry, and Richard Hayes. 2009. A mobile ECG monitoring system with context collection. In *4th European Conference of the International Federation for Medical and Biological Engineering*. Springer, 1222–1225.

[40] Gadgeon Smart Systems Pvt. Ltd. 2020. WebCardio- A wireless, disposable, leadless multi day Holter monitor. Retrieved Jan, 2020 from https://webcardio.in/

[41] Gabriela Marcu, Jakob E Bardram, and Silvia Gabrielli. 2011. A framework for overcoming challenges in designing persuasive monitoring and feedback systems for mental illness. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 1–8.

[42] Manlio F Márquez, Luis Colín, Milton Guevara, Pedro Iturralde, and Antonio G Hermosillo. 2002. Common electrocardiographic artifacts mimicking arrhythmias in ambulatory monitoring. *American heart journal* 144, 2 (2002), 187–197.

[43] Elizabeth A McGlynn, Steven M Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, and Eve A Kerr. 2003. The quality of health care delivered to adults in the United States. *New England journal of medicine* 348, 26 (2003), 2635–2645.

[44] WF McIntyre, Jean Hai Ein Yong, RK Sandhu, DJ Gladstone, K Simek, YY Liu, FR Quinn, R Tytus, D Zizzo, S Henein, et al. 2019. Prevalence of undiagnosed atrial fibrillation in elderly individuals and potential cost-effectiveness of non-invasive ambulatory electrocardiographic screening: The ASSERT-III study. *Journal of Electrocardiology* (2019).

[45] WF McIntyre, Jean Hai Ein Yong, RK Sandhu, DJ Gladstone, K Simek, YY Liu, FR Quinn, R Tytus, D Zizzo, S Henein, et al. 2020. Prevalence of undiagnosed atrial fibrillation in elderly individuals and potential cost-effectiveness of non-invasive ambulatory electrocardiographic screening: The ASSERT-III study. *Journal of electrocardiology* 58 (2020), 56–60.

[46] Fen Miao, Yayu Cheng, Yi He, Qingyun He, and Ye Li. 2015. A wearable context-aware ECG monitoring system integrated with built-in kinematic sensors of the smartphone. *Sensors* 15, 5 (2015), 11465–11484.

[47] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science* 6, 1 (2011), 42.

[48] World Health Organization. 2017. Case study: Type 1 diabetes. Retrieved Aug 8, 2019 from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[49] Patrick X Quesnel, Adrian DC Chan, and Homer Yang. 2014. Signal quality and false myocardial ischemia alarms in ambulatory electrocardiograms. In *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 1–5.

[50] Joel M Raja, Carol Elsakr, Sherif Roman, Brandon Cave, Issa Pour-Ghaz, Amit Nanda, Miguel Maturana, and Rami N Khouzam. 2019. Apple Watch, Wearables, and Heart Rhythm: where do we stand? *Annals of translational medicine* 7, 17 (2019).

[51] Meghan Reading, Dawon Baik, Melissa Beauchemin, Kathleen T Hickey, and Jacqueline A Merrill. 2018. Factors influencing sustained engagement with ECG self-monitoring: perspectives from patients and health care providers. *Applied clinical informatics* 9, 04 (2018), 772–781.

[52] Alireza Sahami Shirazi, Diana Cheng, Oliver Kroell, Dagmar Kern, and Albrecht Schmidt. 2007. CardioViz: Contextual capture and visualization for long-term ecg data. In *adjunct Proceedings of Ubicomp*.

[53] Monica Solbiati, Lucia Trombetta, Roberto M Sacco, Luca Erba, Viviana Bozzano, Giorgio Costantino, Satish R Raj, Franca Barbic, Giovanni Casazza, Franca Dipaola, et al. 2019. A Systematic Review of Noninvasive Electrocardiogram Monitoring Devices for the Evaluation of Suspected Cardiovascular Syncope. *Journal of Medical Devices* 13, 2 (2019), 024001.

[54] F Spadini, F Vergari, L Nachman, C Lamberti, and T Salmon Cinotti. 2008. A wireless and context-aware ECG monitor: An iMote2 based portable system. In *2008 Computers in Cardiology*. IEEE, 997–1000.

[55] Zenicor Medical Systems. 2020. The Zenicor solution. Retrieved July 8, 2020 from https://zenicor.com/zenicor-ekg/

[56] Derek H Tang, Adrienne M Gilligan, and Klaus Romero. 2014. Economic burden and disparities in healthcare resource use among adult patients with cardiac arrhythmia. *Applied health economics and health policy* 12, 1 (2014), 59–71.
[57] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.
[58] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.
[59] Castro Urda Víctor, Parra Esteban Carolina, Toquero Ramos Jorge, Carneado-Ruiz Joaquín, Sánchez García Manuel, Cobo Marcos Marta, Fernández Villanueva José María, Pham Trung Chinh, Ortega Marcos Javier, Mingo Santos Susana, et al. 2018. Incidence and predictive factors of hidden atrial fibrillation detected by implantable loop recorder after an embolic stroke of undetermined source. *Journal of Atrial Fibrillation* 11, 3 (2018).
[60] Edward H Wagner. 1998. Chronic disease management: what will it take to improve care for chronic illness? *Effective clinical practice* 1, 1 (1998).
[61] Peter West, Max Van Kleek, Richard Giordano, Mark J Weal, and Nigel Shadbolt. 2018. Common barriers to the use of patient-generated data across clinical settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

A LIST OF CUMACF QUESTIONS USED IN THE FEASIBILITY EVALUATION

List of CUMACF (CACHET Unified Methodology for Assessment of Clinical Feasibility) Questions used	
	Questions
Q1	Overall, I would find the system useful in home based longitudinal ECG collection for arrhythmia screening
Q2	I would use mCardia daily basis as instructed
Q3	Using mCardia would increases the quality of communication b/w me and my doctor
Q4	Using mCardia would reduce recall bias in reporting my symptoms during screening period
Q5	mCardia would help me in keeping track of my daily activeness and unusual symptoms and help me understand my symptoms better
Effort Expectancy	
Q6	Overall, I would be satisfied with how easy it is to use mCardia App
Q7	My interaction with mCardia would be clear and understandable.
Q8	It would be easy for me to learn to use mCardia App
Q9	I would find mCardia easy to use
Q10	I would be skillful at using mCardia
Q11	The information (such as [error messages | help | messages | guidelines | tutorials | …]) provided with mCardia are clear and useful
Q12	The interface was effective in helping me complete the task [events entry]
Q13	mCardia was pleasant to use
Q14	mCardia has all the functionalities that I expect it to have
Social Influence	
Q15	My doctor thinks that I should use mCardia
Q16	My family [spouse | children | parents | ...] think that I should use mCardia
Facilitating Conditions	
Q17	I have the resources necessary to use mCarida app
Q18	I have the knowledge necessary to use mCardia app
Q19	A specific person should be available for assistance with mCardia if I face any difficulty with mCardia App

A.3 Contextual and Temporal Distribution of False Positives in a Deep Learning-Based Atrial Fibrillation Detection Algorithm: An Investigation

`Authors`
Devender Kumar, Sadasivan Puthusserypady, Helena Dominguez, Kamal Sharma, and Jakob E. Bardram

`Journal`
Expert Systems with Applications. [Under Review]

Contextual and Temporal Distribution of False Positives in a Deep Learning Based Atrial Fibrillation Detection Algorithm: An Investigation

Devender Kumar[a], Sadasivan Puthusserypady[a], Helena Dominguez[b], Kamal Sharma[c], Jakob E. Bardram[a]

[a]*Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.*
[b]*Department of Cardiology,Bispebjerg-Frederiksberg Hospital, Copenhagen, Denmark.*
[c]*U. N. Mehta Institute of Cardiology and Research Centre, Civil Hospital Campus, Ahmedabad, Gujarat, India.*

Abstract

Goal: To investigate the contextual and temporal distribution of false positives (FPs) in a state-of-the-art deep learning (DL)-based atrial fibrillation (AF) detection algorithm when applied to patient-operated ECG from free-living ambulatory conditions. We hypothesize that under such conditions, the FPs detected by a DL model might have some correlations with the patient's ambulatory contexts. *Method*: First, a DL model is trained and evaluated on three public arrhythmia datasets from PhysioNet. It is ensured that the model had state-of-the-art performance on these public datasets. Thereafter, the same model is applied to a 215-days long contextualized single-channel ECG dataset collected under free-living ambulatory conditions. Through a manual examination of the model's output, the ground truth is obtained, and the correlations between patient's ambulatory contexts and the true/false positive rate are analyzed. *Results*: Nearly 62% of the non-trivial short segments of FPs are mainly associated with three specific contexts: change in activity, change in body position (especially during the night), and sudden movement acceleration. Moreover, the number of FPs detected by the DL model are more in female than in male participants. Finally, true positive (TP) AF segments are found more in the morning and late evening. *Significance:* These findings may have significant implications for the current use and future design of DL models for AF detection and help understand the role of context information in reducing the FP rate in real-time AF detection under free-living conditions.

Keywords: Atrial Fibrillation (AF), Electrocardiogram (ECG), Context-aware ECG, Deep Learning (DL), False Positive (FP).

1. Introduction

Cardiovascular diseases (CVD) are one of the most common causes of mortality worldwide, especially in the developed countries (Gaziano, 2001). Among all CVDs, atrial fibrillation (AF) has the highest prevalence and is considered as a major contributor to stroke (Wolf et al., 1987). It affects nearly 10 and 2.3 million people in Europe and the United States, respectively. This is projected to rise to 14-17 million in Europe by year 2030 and 5.6 million

Email addresses: deku@dtu.dk (Devender Kumar), sapu@dtu.dk (Sadasivan Puthusserypady), maria.helena.dominguez.vall-lamora.02@regionh.dk (Helena Dominguez), kamalcardiodoc@gmail.com (Kamal Sharma), jakba@dtu.dk (Jakob E. Bardram)

in the US by year 2050 (Go et al., 2001; Zoni-Berisso et al., 2014). Early diagnosis of AF can absolutely help in preventing heart attacks and subsequent complications significantly.

Electrocardiogram (ECG) analysis is the most common means of AF diagnosis. However, when AF and other arrhythmias are paroxysmal, they might be missed while performing the clinical investigation on short ECG recordings and may remain undiagnosed (Dinakarrao et al., 2019). Diagnosis may thus require longitudinal ambulatory monitoring in the patient's natural settings. Holter monitors and recent advancements in wearable technology have made it possible to continuously collect ambulatory ECG recordings. The challenge, however, is the manual analysis of such long (several hours) ECG recordings by a cardiologist, which is cumbersome, time consuming and hence extremely costly.

To aid with the analysis of longitudinal ECG, a number of computer-aided arrhythmia detection techniques have been developed. They have evolved from traditional feature-engineering (FE)-based techniques to the more recent machine learning (ML) techniques, such as deep learning (DL). Techniques such as wavelet transformation (WT) (Shyu et al., 2004), support vector machines (SVM) (Melgani & Bazi, 2008), hidden Markov models (HMM) (Coast et al., 1990), and frequency domain analysis (Minami et al., 1999) are examples of FE-based approaches. They require domain knowledge and mostly work well on clean ECG data. In contrast, the DL-based techniques do not require manual feature extractions and in recent years, there have been several advances in the application of DL for AF and other types of arrhythmia detection, with a high level of performance (Andersen et al., 2019; Poh et al., 2018; Pourbabaee et al., 2018; Faust et al., 2018; Wang, 2020; Hannun et al., 2019; Xu et al., 2018; Limam & Precioso, 2017; Yao et al., 2017; Chocron et al., 2020).

Despite these promising improvements in AF detection, bringing these algorithms into widespread adoption still remains challenging. When applying these algorithms on longitudinal ambulatory ECG recordings collected in free-living conditions, they result in a high number of non-trivial FP detection and display general performance degradation (Halvaei et al., 2020; Hannun et al., 2019; Ceylan & Özbay, 2007; Yao et al., 2017; Gao et al., 2019). This is primarily because the majority of these algorithms are trained and evaluated on publicly available high quality ECG datasets, which are collected in controlled clinical settings (Parvaneh et al., 2019). Also, in many cases, only a small number of subjects as well as carefully-selected clean-data are used to provide good performance (Fan et al., 2018). The FP rate (FPR) in a longitudinal ambulatory monitoring under free-living conditions, could potentially lead to spurious or over-diagnoses and patient anxiety (Cheung et al., 2018), especially in the population with a lower AF burden.

In this paper, we investigate the relationship between the false positive rate (FPR) of a DL-based AF detection algorithm and the patient's context using a dataset collected in an ambulatory, free-living environment. It is hypothesized that in such conditions, there are specific user context which trigger more FP in a DL-based algorithm. For example, it has been shown that different walking style and pattern influences the FPR; in an abnormal heartbeat detection algorithm more FPs are found when a person walks on a slope rather than on a flat surface (Noh et al., 2013). We explore this more generally and examine the relationship between FPR and patient's context like activities, change in body position (lying supine, lying left/right, standing etc.), movement acceleration, eating

heavy meals, user reported events and unusual symptoms. Understanding the temporal and contextual distribution of FP in free-living condition will help in designing DL models which incorporate such contextual confounders in their design, thereby can achieve context-specific dynamic adjustment of sensitivity and specificity in free-living conditions.

The main contributions of this work are twofold: (i) We identify the user context which result in more FPR in a DL-based model when applied to ECG collected under free-living ambulatory conditions, and (ii) We outline the design implication and the role of context-awareness for the design of future DL-based AF detection models.

The rest of this article is organized as follows: Section 2, describes the related work. The ECG datasets and DL model used in this study are outlined in section 3. In section 4, the experimental details are provided. The obtained results are presented in section 5. Section 6 discusses the results in detail. Limitations and future work are described in section 7, followed by concluding remarks in section 8.

2. Related Work

Over the years, numerous techniques have been developed for computer-aided automatic detection of AF and other types of arrhythmias. This includes algorithms based on auto-correlation function, spectrum analysis, pattern recognition, threshold-crossing intervals, and ML. Although these methods achieve a decent classification accuracy, they face two significant challenges: (1) they require manual feature extraction by the domain experts, and (2) their performance degrades in the presence of noise. To overcome these challenges, researchers have been exploring DL for AF detection. The DL models, such as convolutional neural network (CNN), belief propagation deep neural networks (DNNs), and long-short term memory (LSTM) networks are now being extensively used in AF detection (Dinakarrao et al., 2019). These models help in achieving end-to-end AF detection without any need for manual feature extraction (Pourbabaee et al., 2018; Andersen et al., 2019).

Andersen et al. (2019) built an end-to-end model by combining CNN and recurrent neural network (RNN). In a 5-fold cross-validation on MIT-BIH AF Database (AFDB), the model achieved a sensitivity and specificity of 98.98% and 96.95%, respectively. In addition, the model is claimed to be computationally efficient as it can process 24 hours of ECG in less than a second. Similarly, Petmezas et al. (2021) designed a hybrid CNN-LSTM network that utilizes focal loss and an improved version of cross-entropy loss to improve the AF classification in imbalanced ECG datasets and reported a sensitivity of 97.87% and specificity of 99.29% in a 10-fold cross-validation on AFDB.

Lai et al. (2020) compared four different 8-layer CNN based AF detection models and found that combining two individual models of CNN with F-wave frequency spectrum and RR-interval (RRI) features achieved better performance over the individual model. On a 24 hours long single lead in-hospital ECG, they achieved accuracy, sensitivity, and specificity of 93.1%, 93.1%, and 93.4%, respectively. Similar, multiplicative fusion approach was proposed in `MultiFusionNet` (Tran et al., 2020) in which a network combines two sub-networks. This network architecture is reported to be performing well even with a small training dataset. Besides, Fan et al. (2018) also introduced a multi-scale fusion of deep CNN (MS-CNN) for AF screening from single lead short ECG and achieved a classification accuracy of 96.99% on ECG as short as 5 seconds.

Pourbabaee et al. (2018) introduced a computationally intelligent AF detection model employing deep CNN with raw ECG as inputs. By employing the majority voting methodology for performance improvement,Wei et al. (2019) combined recurrence complex network (RCN) and CNN and have achieved sensitivity, specificity, and accuracy of 94.28%, 94.91%, and 94.59%, respectively. Further, Limam & Precioso (2017) used a convolutional RNN comprising two independent CNN models, where one of them used raw ECG data, and the other used heart rate data as inputs. The features from these two CNNs are then merged into an RNN for AF classification.

Apart from standard ECG, in recent years, other modalities such as photoplethysmography (PPG), contextual and mechanical signals (i.e., seismocardiograms) are also being explored for AF detection (Kwon et al., 2019; Lahdenoja et al., 2017; Hurnanen et al., 2016; Shen et al., 2019). For instance, Hurnanen et al. (2016) used features of the spectral entropy and a heart rate variability index computed from the seismocardiograms and achieved a 99.9% true positive (TP) and 96.4% true negative (TN) AF detection rate. Similarly, Lahdenoja et al. (2017) achieved 97.4% accuracy in healthy versus AF classification using the accelerometer and gyroscope readings of a smartphone.

Although the aforementioned models show good performance on publicly available ECG datasets, their generalisability and performance on ECG collected under free-living conditions remains problematic. For instance, the model by Andersen et al. (2019) resulted in 4.99% FPR when applied on unseen MIT-BIH NSR Database (NSRDB) which only has the data from healthy subjects with normal sinus rhythm (NSR). Similarly, the model by Lai et al. (2020) reported high FPR on ambulatory ECG dataset as compared to its performance on AFDB. Without addressing the FPR problem under the free-living condition, bringing these models into widespread use will not be possible as FPR (as small as 1%) could lead to over-diagnosis and patient anxiety in a longitudinal screening (Cheung et al., 2018; Komorowski & Celi, 2017).

Moreover, there is an increasing awareness that the user's activity and context is relevant in understanding and analyzing ambulatory ECG (Ebrahimi et al., 2020; Dinakarrao et al., 2019). For example, it has been shown that a heartbeat classification algorithm based on ECG has more FPs when the subject is walking on a sloped surface as compared to walking on a flat surface or sitting (Noh et al., 2013). In general, it is more and more acknowledged that under free-living conditions, the occurrence of FPR in an AF classification algorithm have some correlations with the user's context. Thus, without understanding the patient's context, ECG analysis under free-living conditions remains prone to a high FPR, which again may lead to under/over diagnosis.

3. Research Methods

To investigate the contextual and temporal relationships of FPs in a DL model applied on ambulatory data, the following approach is followed. First, an end-to-end AF detection DL model is trained and ensured that it had the state-of-the-art performance on publicly available ECG datasets. This model is then tested on a ambulatory and contextualized ECG dataset collected from patients under free-living conditions. Thereafter, the AF onset and offset detected by the model are manually examined and labeled by biomedical engineers and cardiologists. Finally, the patient's context is plotted against the FP and TP episodes detected by the model.

3.1. Public arrhythmia datasets

This research is based on three public datasets from PhysioNet (Goldberger et al., 2000): the MIT-BIH AF Database (AFDB) (Moody, 1983), the MIT-BIH Arrhythmia Database (MITDB) (Moody & Mark, 2001), and the MIT-BIH NSR Database (NSRDB) (Goldberger et al., 2000). Table 1 lists the technical specifications of these databases. These have labeling of different types of arrhythmias. As this work focuses on AF, the data is categorized into two classes: "AF" and "other" (which includes noise, NSR, and other types of arrhythmia (if any) combined).

Table 1: Technical specifications of databases. Ch: No. of ECG channels, T: Duration of the data recording, F: Sampling frequency, N: No. of unique subjects in the recording, T-AF: Duration of AF, R: No. of unique annotated rhythms

Databases	**Ch**	**F (Hz)**	**T (h)**	**T-AF (h/%)**	**R**	**N**
AFDB	2	250	234.3	93.40 (39.87%)	4	25
MITDB	2	360	24.07	2.16 (8.97%)	15	47
NSRDB	2	128	437.5	0 (0%)	1	18

3.2. Contextualised ECG dataset

None of the public arrhythmia databases from PhysioNet (Goldberger et al., 2000) have any information about the patient's context during the ECG recording. Moreover, despite being ambulatory, ECG recordings are limited to a few hours as well as activities. Therefore, they do not contain ECG morphology changes and noise contamination under all activities (e.g., biking, climbing stairs, running, jogging, walking) that occur in natural free-living settings.

In this study, the mCardia system (Kumar et al., 2021) is used for collecting the longitudinal single channel contextualized ECG in the patient's natural setting. The resulting dataset is named as the "CACHET Contextualised Arrhythmia Database (CACHET-CADB)"(CACHET, 2021). The participants are recruited from Denmark and India. The ethical approval for the data collection is obtained from the Institutional Review Board of the Mahatma Gandhi Medical College and Research Institute, Jaipur, India, and the Danish Research Ethical Committee. Patients' recruitment took place during their outpatient arrhythmia clinic visit and only patients who are either already diagnosed with paroxysmal AF or suspected of having AF are included. The data collected from 21 patients (13 male and 8 female) with an average age of 58.7±10.11 years is used in this study. The length of ECG recordings from patients are varied from 3 days to 3 weeks.

Table 2 lists the details of the collected data. The mCardia system uses a Movisens single-channel ECG EcgMove4 monitor mounted on the chest of the patient, which samples ECG at 1024Hz with a 12-bit resolution (Movisens GmbH, 2019). Contextual data is collected both from the EcgMove4 device and from the participant's mobile phone. This data can be categorized into three main types: (i) passively sensed data such as activity, acceleration, body position, noise, and local weather; (ii) patient-reported data such as sleep quality, stress level, food consumption; and (iii) experienced symptoms recorded as 'events', such as palpitations or shortness of breath. For the event-based data, participants could tap on the EcgMove4 device if they felt any unusual symptoms and then provide additional information in the mCardia app, which included specifying the duration of the symptoms and the activity they

Table 2: ECG and contextual data types in the CACHET-CADB with source and sampling rate. S: Sensed. PR: Patient-reported. EB: Event-based.

Data type	**Type**	**Data source**	**Sampling rate**
ECG	S	EcgMove4	1024 Hz
HR	S	EcgMove4	1/60Hz
Movement acc.	S	EcgMove4	64Hz
Rotation rate	S	EcgMove4	64Hz
Body position	S	EcgMove4	1/60Hz
Activity	S	Phone	EB
Steps	S	EcgMove4 & Phone	1/60Hz & EB
Events	PR	EcgMove4 & Phone	EB
Weather	S	Phone	4/day
Sleep	PR & S	Phone	1/Day
Noise level	S	Phone	1/120 Hz
Dietary	PR	Phone	1/Day

are doing when the event occurred. More fine grind (every 10 seconds) activity, movement acceleration, and body position data are obtained by processing the raw 3D accelerometer and gyroscope data from the EcgMove4 device using the Movisens DataAnalyzer tool (Movisens GmbH, 2020). Patient-reported data, such as symptoms events, food intake and stress levels, are used without pre-processing.

3.3. Model overview

An end-to-end AF detection model, which is a modification of Andersen et al. (2019) for a single channel ECG is developed for this study. The original model is trained and tested on two channel ECG from AFDB dataset, but since the CACHET-CADB uses only a single channel, we choose to modify and train the model on a single channel ECG. The multi-layer CNN is utilized for extracting high-level features from the raw input sequence and RNN is used for processing the sequential features extracted by CNN. Figure 1 shows the high-level architecture of this DL model.

3.3.1. Data preparation and pre-processing

To train the model on the AFDB dataset, we utilize the RRI as input for feature extraction. Irregular ventricular contraction due to AF is reflected in the RRIs and have been used in several studies for efficient detection of AF (Andersen et al., 2019; Zhou et al., 2014; Nguyen et al., 2018; Moody, 1983). Moreover, it is computationally less expensive and thus can be utilized for real-time AF detection in wearables. The public databases (AFDB, MITDB, and NSRDB) already provide the locations of R-peaks in the raw ECG, which is used to find the RRI

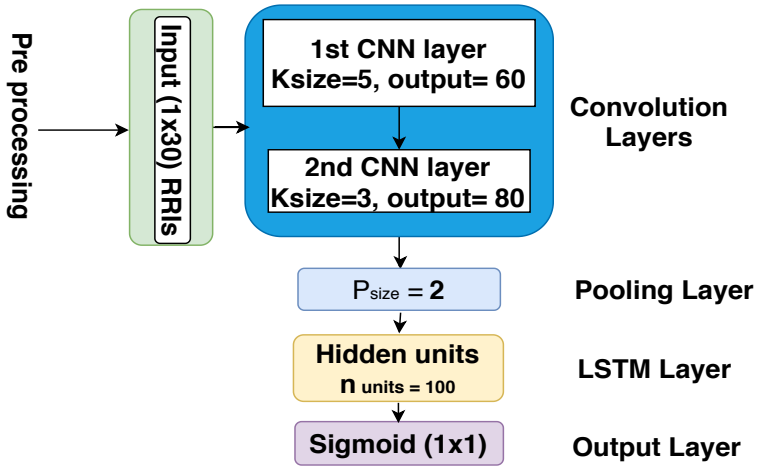


Figure 1: The network consisted of 2 convolutional layers, followed by a pooling layer. Features extracted in the convolution are fed into the LSTM layer, which consists of 100 hidden units. Finally, the sigmoid layer gives the binary output of the entire 30 RRIs long window. The network takes 30 RRIs (calculated from raw ECG) data as input and outputs a binary classification of AF or non-AF class.

using the following formula:

$$RRI(n) = \frac{rPeaks(n+1) - rPeaks(n)}{f_s}, \tag{1}$$

where f_s is the sampling rate. These RRIs are segmented into windows of 30 RRIs and fed as inputs to the model. Each consecutive window has an overlap of 10 RRIs. The beat-by-beat annotations are used to create the new annotation of the entire window of 30 RRIs. If the fraction of AF beats in the windows' length is more than 0.5, then the entire window is labeled as AF (Andersen et al., 2019).

The length of ECG recordings in the CACHET-CADB varies from a minimum of three days to three weeks and are therefore trimmed into segments of 24 hours to ease the processing and analysis. Since it is an ambulatory ECG dataset and contains many artifacts, we used a set of filters and techniques to minimize the artifacts: (i) Band-pass [0.5 to 50 Hz] filter to remove artifacts and baseline wander, (ii) Savitzky-Golay filter (Press & Teukolsky, 1990) for smoothing the data, and (iii) 10-seconds sliding windows to calculate the cross-correlation and reject noisy signal. Thereafter the Pan-Tomkins algorithm (Pan & Tompkins, 1985) is used to calculate the R-peaks from the ECG signal. From these R-peaks, the RRIs are obtained using Eq. (1). It is important to note that the CACHET-CADB dataset is only used for testing the hypothesis; model is trained on the AFDB.

3.3.2. Model architecture

The processing steps of the model is illustrated in Figure 1. After pre-processing, the windows of 30 RRIs are fed into the convolution layer, which extracts the temporal features from the input signal. The model consists of two successive convolution layers. The first layer has a kernel size (K_{size}) = 5, and its output filter (n_{filter}) consists

of 60 learned features. The second convolution layer has these 60 features as its input with a kernel size (K_{size}) = 3 and n_{filter} = 80 output features. In both CNN layers, the input is zero-padded to preserve the temporal dimensions.

After that, the max-pooling operation is performed to retain maximum spatial information while reducing the temporal dimensions of the extracted features in the convolution stage. The pooling layer uses a kernel size of $P_{size} = 2$ with strides of two and reduces the temporal dimension by half. The pooling layer output is passed as input to the bidirectional LSTM layer. It consists of $n_{units} = 100$ hidden units. Finally, the output of the LSTM layer is passed to a single sigmoid neuron. It outputs a probability of a given input sequence of 30 RRIs belonging to AF. The default probability threshold is set to 0.5. If the probability is ≥ 0.5, the input sequence belongs to the AF class; otherwise, it is classified as "other class". Further details of the model and its tuning parameters can be found in Andersen et al. (2019).

3.4. Model evaluation metrics

In order to evaluate the performance of the AF detection model on the different arrhythmia datasets, we applied the sensitivity (Se), specificity (Sp), FPR, and accuracy (Acc), which are the standard metrics used for evaluating the performance. They are defined as follows:

$$Se = \frac{TP}{TP + FN}, \tag{2}$$

$$Sp = \frac{TN}{TN + FP}, \tag{3}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$FPR = \frac{FP}{FP + TN}. \tag{5}$$

where TP, FN, and TN stands for true positives, false negatives, and true negatives, respectively.

4. Experiment

Figure 2 shows the workflow of our experiment. It involves the following four steps: (A) Testing the DL model on public datasets and comparing its performance with state-of-the-art on these datasets, (B) Applying the model on the CACHET-CADB, (C) Manual annotation of the model's output for the ground truth, and (D) Mapping the contextual data to the TP and FP predicted by the model. Each of these steps are detailed below.

4.1. Testing model on public datasets

In the first step, we train and test the model's performance on the public databases from PhysioNet (AFDB, MITDB, NSRDB) and compare it with state-of-the-art performance by other DL models on these databases. It should be noted that compared to other models in the literature, which are mostly trained and tested on two-channel

Table 3: Performance of the model on public datasets and its comparison with other state-of-the-art models. C: No. of ECG channels

Algorithm	**C**	**AFDB**				**MITDB**				**NSRDB**			
		Se (%)	Sp (%)	Acc (%)	FPR (%)	Se (%)	Sp (%)	Acc (%)	FPR (%)	Se (%)	Sp (%)	Acc (%)	FPR (%)
Xia et al. (2018)	1	98.79	97.87	98.63	-	-	-	-	-	-	-	-	-
Lai et al. (2019)	1	97.4	97.2	97.3	-	-	-	-	-	-	-	-	-
Wei et al. (2019)	2	94.28	94.91	94.59	-	-	-	-	-	-	-	-	-
Yao et al. (2017)	2	98.22	98.11	98.18	-	-	-	-	-	-	-	-	-
Dang et al. (2019)	2	99.93	97.03	96.59	-	-	-	-	-	-	-	-	-
Andersen et al. (2019)	2	98.17	96.29	97.10	3.71	98.96	86.04	87.40	13.96	-	95.01	-	4.99
This work	1	96.06	98.29	97.04	1.7	96.87	86.94	87.98	13.06	-	94.44	-	5.56

ECG of these public datasets, our model for this experiment is trained on single-channel (channel-1 of AFDB). It is primarily because our CACHET-CADB used for the contextual analysis contains only single-channel ECG.

4.2. Applying model on CACHET-CADB and mapping model output to raw ECG timestamp

After achieving state-of-the-art performance on the public datasets, the trained model is applied to the CACHET-CADB. Since the input to the DL model are RRIs (after classification), they are mapped back to raw ECG (see Figure 3), and the timestamp of each AF onsets and offsets marked by the DL model are stored in a CSV file. The duration between a single AF onset/offset pair will be referred as a "*segment*" throughout the rest of this article.

One biomedical engineer and a cardiologist manually screened the CACHET-CADB and categorized each participant's recording into one of the three types: (i) NSR (T1), (ii) paroxysmal AF (T2), and (iii) persistent AF (T3). Although, the patients who are either previously diagnosed with AF or suspected of at high risk for AF are recruited for this study, it is observed during the manual screening of the database that a few subjects had persistent AF, i.e., subjects remain in AF for 99% of the total recording time. These subjects are excluded from the study, since contextual analysis of FP/TP in such subjects is not useful for testing our hypothesis.

4.3. Annotation

As illustrated in Figure 2, only type T1 and T2 recordings are manually annotated. A two-stage screening strategy is adopted. In the first stage, each AF onset and offset segments marked by the DL model are reviewed by a biomedical engineer. The obvious noisy or NSR segments are labeled in this round. The ambiguous segments are passed into the second stage, where two independent cardiologists annotated each segment. The two cardiologists' annotations are compared, and any disagreements are resolved to obtain the final labels. Based on this approach, each segment of AF marked by the DL model is assigned a "ground truth" label as either TP or FP.

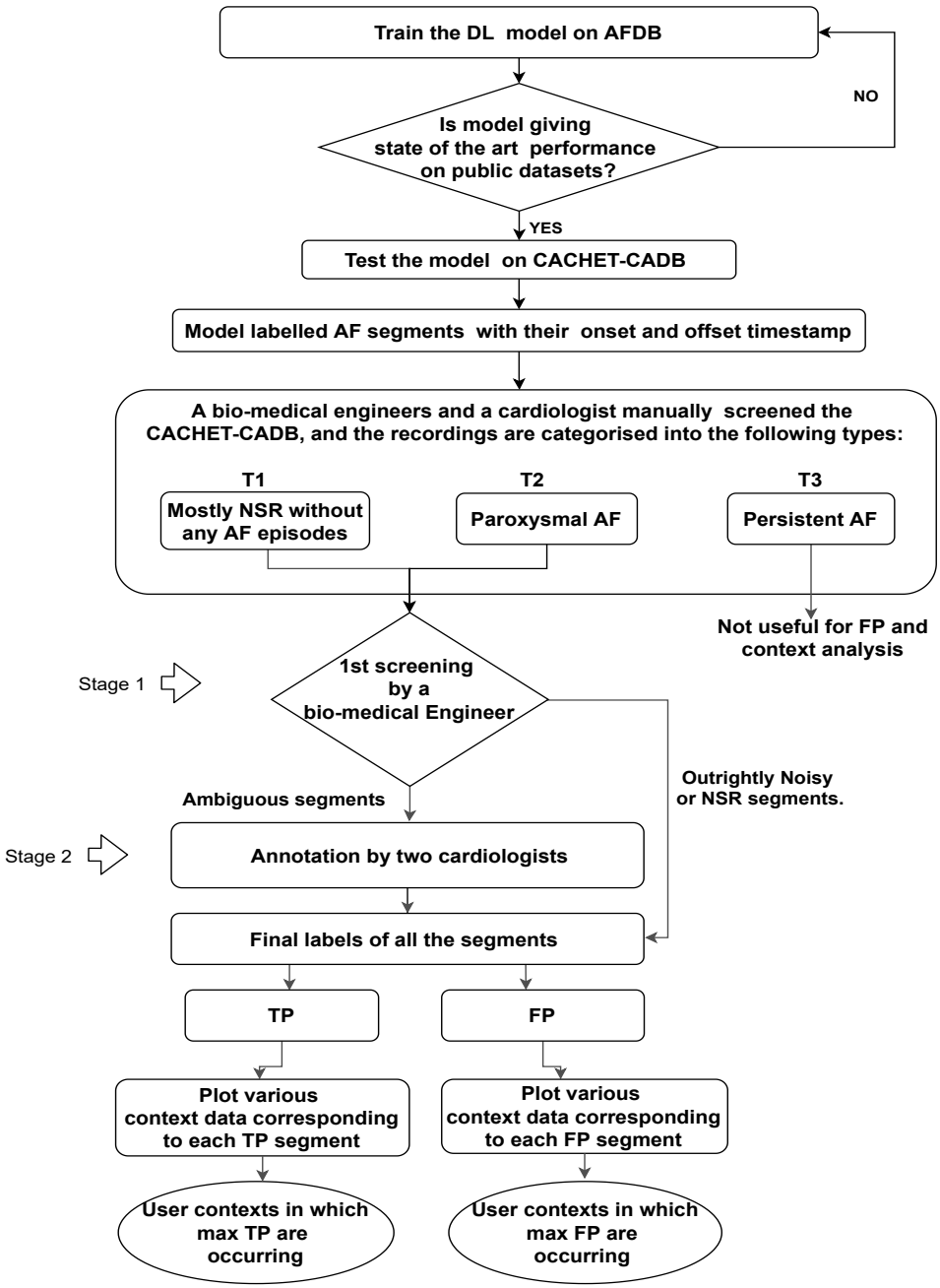


Figure 2: Experimental workflow.

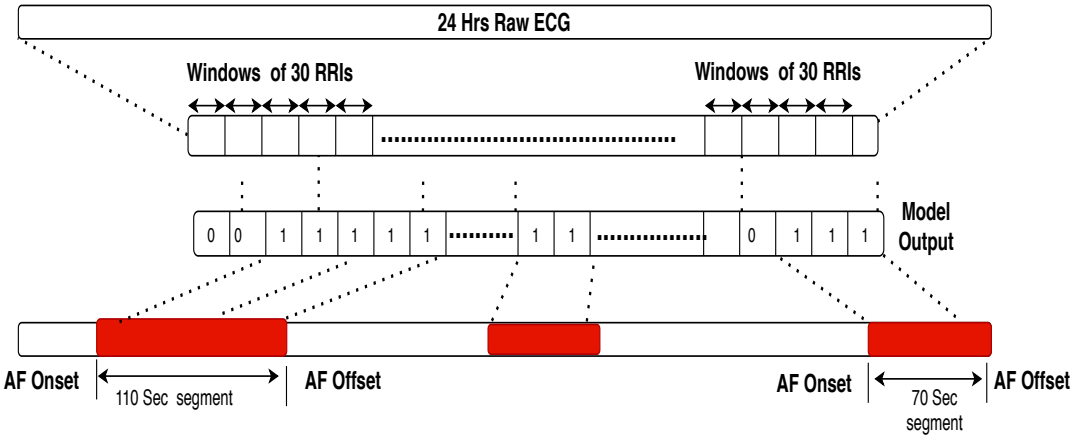


Figure 3: Process of mapping DL model's output to raw ECG timestamp. The AF segments between onset and offset are marked in red.

4.4. Mapping contextual data to TP and FP

After the manual annotation process, the patient's context (movement acceleration, activity, body position, tap events, reported stress level, food intake, etc.) corresponding to those TP and FP labels are analyzed.

5. Results

The DL model for AF detection used in this work is built in Python 3.7 using Keras and Tensorflow 2.0. The model utilizes RRI for feature extraction instead of the raw ECG; thus, the computational time and resource requirement is less compared to the conventional DL models. On a Dual-Core Intel Core i7 (3.6 GHz) CPU with 16 GB RAM and Intel Iris Plus Graphics 650 graphic card computing environment, it takes 0.62 seconds to classify 24 hours of ECG and 2.65 seconds to map the model output to ECG.

5.1. Performance on the PhysioNet datasets

Table 3 lists the performance of the model on three public datasets from PhysioNet. In a 5-fold cross-validation, the Se, Sp, Acc, and FPR on AFDB are shown to be 96.06%, 98.29%, 97.04%, and 1.7%, respectively. We trained and tested on ECG channel-1 and channel-2 of the AFDB independently, and found the performance on channel-1 slightly better than channel-2. Thus we chose to train the model on channel-1 for the rest of the analysis. Both the MIT-BIH Arrhythmia Database (MITDB) and NSRDB are used only for testing the model's generalization and, as expected, the FPR on these unseen databases is higher (13.06% and 5.56%, respectively).

5.2. Analyzing FPs in the CACHET-CADB

A total of 215 days single-channel ECG of 21 subjects from the CACHET-CADB is analyzed using the DL model. During the manual screening, four subjects are found to remain in persistent AF and are excluded from further analysis (see Figure 2). Table 4 shows the total number of segments marked as AF by the DL model, the

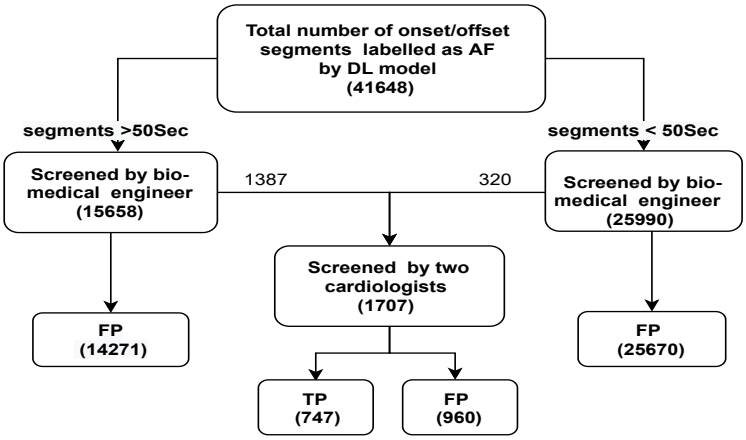


Figure 4: Annotation summary after manual examination

number of recording days, the average number of onset/offset segments per day, and the number of segments of length ≤ 50 seconds and ≥ 400 seconds. An average 261 segments per day are flagged as AF by the DL model of which more than 62% are of length ≤ 50 seconds.

Figure 4 shows the results of the manual annotation process. A total of 41,648 segments are marked as AF by the DL model. A biomedical engineer, in the first stage, manually annotated and looked for obvious noisy/NSR segments. When in doubt, a segment is sent to the second stage and annotated by two cardiologist. In the second stage, a total of 1,707 segments are annotated, of which 747 turned out to be TP, and the remaining 960 are FP.

5.2.1. Correlations between FPs and context features in the CACHET-CADB

As shown in Table 4, nearly 62% of the total AF segments detected by the model are of length less than 50 seconds, and 99.9% of them are FP (Figure 4). Through visual inspection, it is found that these short segments are mostly associated with a change in body position, movement acceleration, and activity change. Figure 5 shows 24 hours of ECG with the AF segments detected by the DL model, the ground truth (true labels after annotation by the cardiologists), and their correlations with the user-context. Whenever there is a sudden peak in movement acceleration, and if the DL model has classified that ECG segment as AF, it is mostly FP when it is of length less than 50 seconds. In contrast to these many FP from short segments, it is observed that long segments of FP are contributed by multiple premature ventricular contraction (PVC) beats (bigeminy, trigeminy etc.) and atrial flutter.

When investigating the correlation between TP and user-generated reports on stress and food intake, no conclusive pattern is found. This is primarily due to the fact that many of the patients in paroxysmal AF failed to fill out the stress and food surveys regularly, and those who filled them regularly are either in continuous AF or did not have any AF episodes at all.

With respect to patient reported tap marker events and time in general, the analysis of TP is done for only

Table 4: Statistics of DL model's performance on CACHET-CADB. AF-DL: Number of segments detected as AF by DL model

Subject	AF-DL	Days	Avg/day	Seg≤50s	Seg≥400s
S1	1991	12	166	1348	68
S2	2826	5	565	1857	25
S3	3419	16	214	1705	317
S4	3712	10	371	1711	107
S5	2308	11	209	1381	166
S6	3198	12	266	1877	65
S7	1702	12	141	1374	32
S8	3058	8	382	2132	153
S9	2415	12	201	1646	55
S10	4290	16	268	2835	71
S11	1236	19	65	883	37
S12	2470	12	205	1707	116
S13	1453	14	103	764	283
S14	787	5	157	614	8
S15	2075	4	518	1152	46
S16	2742	8	342	1569	97
S17	1966	7	280	1435	10

subjects in paroxysmal AF (T2 in Figure 2). Three subjects from this category in our dataset are found to have more TP AF onsets/offsets clustered around morning and late evening hours. Also, most of the self-reported tap marking events of unusual symptoms were also around these time periods. The most common symptom reported in these self-reported events are "shortness of breath" and "palpitation".

Finally, an investigation on the possible gender differences in the number of FP detected by the model showed that the female participants are more prone for FP than male participants. This could be attributed to the mounting of the ECG device on the chest having breast movements for females adding more confounding motion artifacts. Besides, as ECG patterns are significantly different between genders (Surawicz & Parikh, 2002); this result could also be due to the bias of female gender distribution in the AFDB (training) dataset.

6. Discussion

The DL model designed and trained in this paper achieved state-of-the-art performance when applied to the AFDB. However, the number of FP are increased when applied on the MITDB and NSRDB, and, as hypothesised, the number of FP increased even more when applied on the CACHET-CADB, which contains data collected under free-living ambulatory conditions. In this section, a discussion of these findings and reflection upon their implications

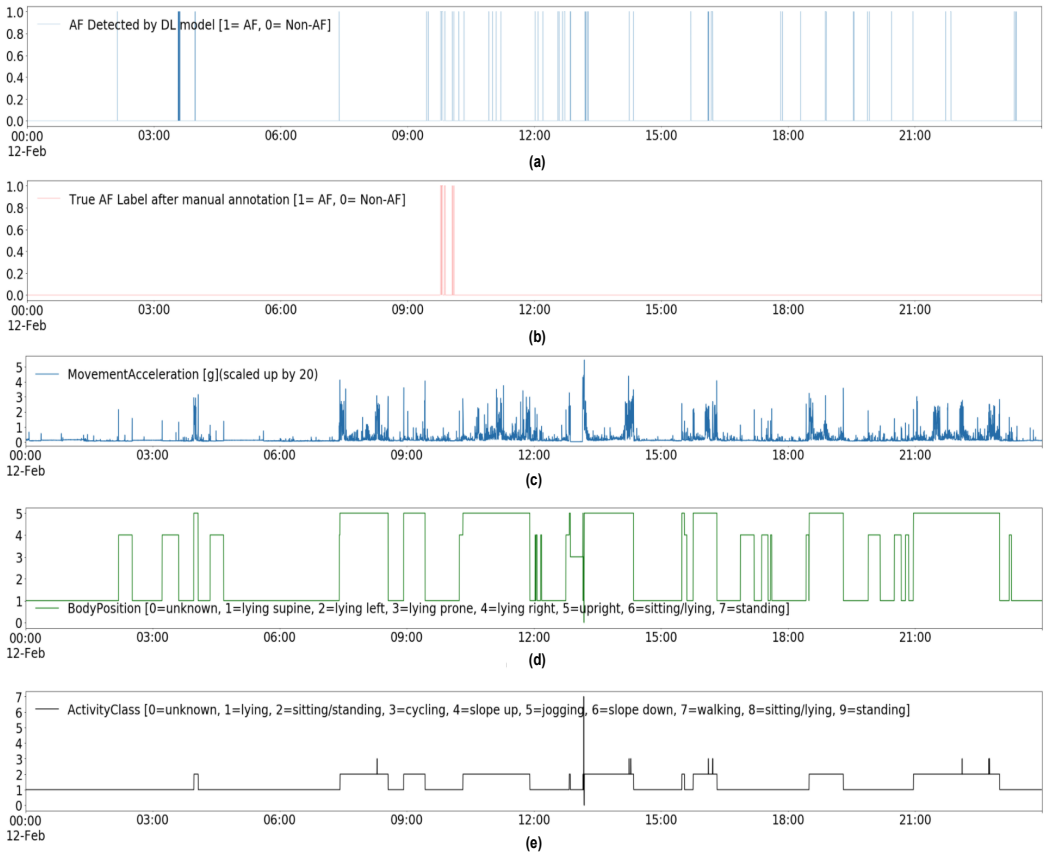


Figure 5: User context and FP occurrences in a 24 hour ECG. The AF detected by the DL model in 24 hours ECG is shown in (a). The ground truth of AF episodes after manual annotation is shown in (b). The short segments ($\leq$ 50-seconds) of FP detected by the DL model in (a) are associated with movement accelerations peaks in (c) and the body position and activity change in (d) and (e).

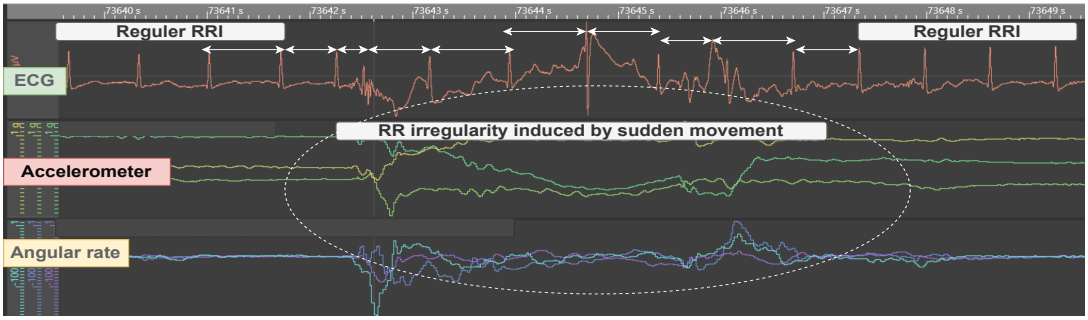


Figure 6: Movement acceleration induced irregularity in RRIs resulting in FP detection.

on the design of AF detection models and their use for CVD monitoring and treatment is provided.

6.1. False positives and user context

Our results show that despite showing good performance on public datasets (Table 3), applying the DL model on ambulatory ECG recordings collected under free-living conditions resulted in a very large number of non-trivial FP. When investigating the correlation between the FP and the user's context, it is revealed that the majority of short length non-trivial FP are associated with three primary context features; (i) activity change, (ii) body position change (especially during the night), and (iii) movement acceleration (see Figure 5).

Based on these findings, the DL model could be significantly improved by taking such contextual information into consideration. If a segment marked as an AF onset is at the start of user activity change (i.e., sitting to walking or running), or a body position change during the night (from supine to lying on the left side), and is not lasting for more than 50 seconds, then there is a high probability that this is a FP. Figure 6 depicts how the change in context induces the variability in RRI, which resembles the ECG during AF. A similar pattern can also be seen with the movement acceleration in Figure 5; whenever there is a peak in movement acceleration and if the DL model has marked an AF onset on this peak, it is mostly a FP if it last less than 50 seconds.

As for the temporal distribution of AF, it is well-known that AF episodes are more prevalent early in the morning and late in the evening (Viskin et al., 1999; Hansson et al., 2004). In our analysis, it is found that a small number of subjects who had TP AF occurrences in the morning and evening. The knowledge about such temporal patterns to AF episodes can be utilized to dynamically fine-tune the sensitivity and specificity of an AF detection algorithms during such AF-prone time periods.

Since it is known that mental stress and certain types of food and drink intake (e.g., alcohol, coffee or a heavy meal) are common triggers of paroxysmal AF (Hansson et al., 2004), it is important to consider these factors. However, due to the patients' lack of compliance in providing this information regularly during the data collection period, it is not possible to see any such conclusive patterns in the analysis. To convey this, the availability of automatic bio-markers for detecting food intake (Sazonov & Fontana, 2011) and mental stress (Hovsepian et al., 2015) may be useful in the future for dynamic adjustment of the DL models output.

6.2. Implications for algorithm design

The results from this study clearly show the influence of the patient's contexts on the algorithm's FPR. This insight gives us the following options that can improve the AF detection algorithm's performance.

Firstly, as shown in this study, nearly 99.9% of the short segment ($\leq$ 50 seconds) are falsely marked as AF when there is a change in activity, body position or movement acceleration. A context-aware heuristic can be built to adjust the probability of sigmoid function. For instance, when an AF is detected around a change in activity or body position, such segments can be put in the buffer, and subsequent segments can be observed before deciding the final probability. Context information can also be utilized with methods such as a majority voting scheme (Wei et al., 2019; Hurnanen et al., 2016) for improving the model's performance.

Similarly, falls are common amongst elderly patients suspected of AF (Hung et al., 2013), and Se and Sp of algorithms can be fine-tuned around such context information. For example, for a given 30 seconds of input ECG sample, the DL model gives the probabilities as 0.4 or 0.49, whereas the cutoff probability for classifying a sample as AF is set as ≥ 0.5. However, if the given sample belonging to an elderly patient, and a fall has been detected (e.g., via other contextual sensors like the accelerometer) just before this sampling window, the final probability should factor the fall and increase the likelihood of this sample being a TP despite the model's probability slightly less than 0.5.

Secondly, since most ($\approx 60\%$) of the total FP are of a very short length ($\leq$50 seconds) and happening in specific contexts such as changing body position in the bed, on activity change, sudden movement acceleration implies that ECG morphology corresponding to these activities is missing in the public datasets. Taking inference from these findings, future DL models should try to include additional ECG data specific to these contexts in ambulatory free-living conditions. Combining additional ECG morphology in training sets (in addition to public datasets) from these specific contexts would help reduce the FPR.

Thirdly, it has been argued that designing models capable of utilizing multi-modal data can improve the limitations posed by models trained on any single modality (ECG alone) (Hong et al., 2020). Although several researchers have explored techniques like ECG signal quality indexing and multi-modal signals to reduce false alarms, this work is limited to a clinical settings (Aboukhalil et al., 2008; Sadr et al., 2016; Behar et al., 2013). It is not possible to get many of these multi-modal signals from wearables when doing AF monitoring under free-living conditions. However, in the ambulatory setting, contextual features extracted from other sensors on the ECG device in combination with data collection from mobile phones (both sensors and user-generated reports) can be utilized as input signals for the design of multi-modal DL algorithms. For example, accelerometer and gyroscope data has been shown useful for AF assessment (Lahdenoja et al., 2017). Hence, utilizing such contextual data alongside ECG for a multi-modal algorithm design can help in improving the classification of AF under free-living conditions.

6.3. Challenges in bringing mHealth based longitudinal AF screening under free-living conditions

The higher FPRs and incorrect diagnosis in these DL models under free living condition would require careful manual check if used in ambulatory monitoring of patients. This clearly would increase the cost of ambulatory monitoring and is not practically feasible. The risk of inadequate AF diagnosis and treatment in the form of anti-coagulation may put patients at risk of bleeding complications (Kirley et al., 2012; Sørensen et al., 2013; Carley et al., 2014). Moreover, if this type of automatic AF detection algorithms is build into patient-facing mobile health (mHealth) technology, false diagnosis of AF may lead to anxiety for the patients and their relatives. Algorithms for ambulatory, real-world monitoring and diagnosis of AF needs to be realigned to the real-time, longitudinal, contextual, and 'messy' nature of patients' free-living conditions. And importantly, such algorithms must be built and evaluated on multi-site ECG data collected from free-living conditions rather than clinically prepared 'clean' datasets on which they tend to give high accuracy.

7. Limitations and Future Work

The DL model used in this study is trained using RRI features. Usually, the performance of the model trained on RRI features tends to degrade in the presence of multiple PVCs. To deal with PVCs related FPs, an ECG delineation model with a p-wave count can be combined with this model. Also, as described in the experiment section, due to resource limitation, the first round of screening for removing the noisy or NSR segments is done by a biomedical engineer and not by cardiologists. Although the biomedical engineer has the expertise to differentiate noise/NSR rhythms and is advised to pass on all doubtful segments to the cardiologists, there still might be a possibility of mis-classification, especially segments where the AF and noise mimic each other. Moreover, the cardiologists only looked at ECG samples from the onset, offset, and a few random samples between onset and offset when annotating the whole segment. There is a possibility, especially in long AF segments, that a few short ECG snippets in-between are mis-classified as AF and would go unnoticed. Nevertheless, this strategy is very practical for screening the longitudinal ECG data needed for this experiment.

Based on the findings from this experiment, in the future, we would like to build a post-processing heuristic based on contextual features pertaining to movement acceleration, body position, and activity changes that can significantly reduce the FPR in longitudinal monitoring under free-living conditions.

8. Conclusion

In this paper, we explored the contextual and temporal distributions of FPs in a DL algorithm when applied on contextualised ECG collected under free-living conditions. As hypothesized, the algorithm, which has state-of-the-art performance on public datasets, resulted in a large number of FPs when applied on 215-days long patient-operated ambulatory ECG data. Upon analyzing the FPs and the users' context, we found that nearly 99.9% of segments of length $\leq$50 seconds which are falsely labeled as AF are associated with three user contexts, namely, (i) activity change, (ii) change in body position (especially during the night), and (iii) on sudden movement acceleration. Besides, the number of FPs is relatively more in female subjects. The user's self-reported "events" of unusual symptoms during the ECG recording period and the TP segments are clustered in the morning and late evening hours. The paper also discussed the design implications of these findings for future DL models and how contextual features can be utilized for reducing the FPR of AF detection models, when used in free-living ambulatory conditions. We believe that understanding the DL-based end-to-end AF detection models' outcome in the patient's ambulatory context can bring transparency and help identify the sources of the algorithm's shortcomings that otherwise remain a black box.

In recent years, we are witnessing an increasing proliferation of patient-facing mHealth technologies that incorporate such AF detection models. To ensure trust in the wider adoption of such technologies, we find that it is essential to address the FPR in AF detection models.

Acknowledgment

This work was supported in part by the Innovation fund Denmark under grant # 6153-00009B (REAFEL).

References

Aboukhalil, A., Nielsen, L., Saeed, M., Mark, R. G., & Clifford, G. D. (2008). Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of biomedical informatics*, *41*, 442–451.

Andersen, R. S., Peimankar, A., & Puthusserypady, S. (2019). A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications*, *115*, 465–473.

Behar, J., Oster, J., Li, Q., & Clifford, G. D. (2013). Ecg signal quality during arrhythmia and its application to false alarm reduction. *IEEE transactions on biomedical engineering*, *60*, 1660–1666.

CACHET (2021). The CACHET Contextual Arrhythmia Database - CACHET-CADB. URL: https://www.cachet.dk/resources/data/cachet-cadb.

Carley, B., Griesbach, S., Larson, T., & Krueger, K. (2014). Assessment of dabigatran utilization and prescribing patterns for atrial fibrillation in a physician group practice setting. *The American Journal of Cardiology*, *113*, 650–654.

Ceylan, R., & Özbay, Y. (2007). Comparison of fcm, pca and wt techniques for classification ecg arrhythmias using artificial neural network. *Expert Systems with Applications*, *33*, 286–295.

Cheung, C. C., Krahn, A. D., & Andrade, J. G. (2018). The emerging role of wearable technologies in detection of arrhythmia. *Canadian Journal of Cardiology*, *34*, 1083–1087.

Chocron, A., Oster, J., Biton, S., Mendel, F., Elbaz, M., Zeevi, Y., & Behar, J. (2020). Remote atrial fibrillation burden estimation using deep recurrent neural network. *IEEE Transactions on Biomedical Engineering*, .

Coast, D. A., Stern, R. M., Cano, G. G., & Briller, S. A. (1990). An approach to cardiac arrhythmia analysis using hidden markov models. *IEEE Transactions on biomedical Engineering*, *37*, 826–836.

Dang, H., Sun, M., Zhang, G., Qi, X., Zhou, X., & Chang, Q. (2019). A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals. *IEEE Access*, *7*, 75577–75590.

Dinakarrao, S. M. P., Jantsch, A., & Shafique, M. (2019). Computer-aided arrhythmia diagnosis with bio-signal processing: A survey of trends and techniques. *ACM Computing Surveys (CSUR)*, *52*, 1–37.

Ebrahimi, Z., Loni, M., Daneshtalab, M., & Gharehbaghi, A. (2020). A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, (p. 100033).

Fan, X., Yao, Q., Cai, Y., Miao, F., Sun, F., & Li, Y. (2018). Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ecg recordings. *IEEE journal of biomedical and health informatics*, *22*, 1744–1753.

Faust, O., Shenfield, A., Kareem, M., San, T. R., Fujita, H., & Acharya, U. R. (2018). Automated detection of atrial fibrillation using long short-term memory network with rr interval signals. *Computers in biology and medicine*, *102*, 327–335.

Gao, H., Liu, C., Wang, X., Zhao, L., Shen, Q., Ng, E., & Li, J. (2019). An open-access ecg database for algorithm evaluation of qrs detection and heart rate estimation. *Journal of Medical Imaging and Health Informatics*, *9*, 1853–1858.

Gaziano, J. M. (2001). General considerations of cardiovascular disease. *Braunwald's Heart Disease: a Textbook of Cardiovascular Medicine*, *7*, 1–19.

Go, A. S., Hylek, E. M., Phillips, K. A., Chang, Y., Henault, L. E., Selby, J. V., & Singer, D. E. (2001). Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the anticoagulation and risk factors in atrial fibrillation (atria) study. *Jama*, *285*, 2370–2375.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, *101*, e215–e220.

Halvaei, H., Svennberg, E., Sörnmo, L., & Stridh, M. (2020). False alarm reduction in atrial fibrillation screening. In *2020 Computing in Cardiology* (pp. 1–4). IEEE.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, *25*, 65.

Hansson, A., Madsen-Härdig, B., & Olsson, S. B. (2004). Arrhythmia-provoking factors and symptoms at the onset of paroxysmal atrial fibrillation: a study based on interviews with 100 patients seeking hospital assistance. *BMC cardiovascular disorders*, *4*, 13.

Hong, S., Zhou, Y., Shang, J., Xiao, C., & Sun, J. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, (p. 103801).

Hovsepian, K., Al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., & Kumar, S. (2015). cstress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 493–504).

Hung, C.-Y., Wu, T.-J., Wang, K.-Y., Huang, J.-L., Loh, E.-W., Chen, Y.-M., Lin, C.-S., Lin, C.-H., Chen, D.-Y., & Tang, Y.-J. (2013). Falls and atrial fibrillation in elderly patients. *Acta Cardiologica Sinica*, *29*, 436.

Hurnanen, T., Lehtonen, E., Tadi, M. J., Kuusela, T., Kiviniemi, T., Saraste, A., Vasankari, T., Airaksinen, J., Koivisto, T., & Pänkäälä, M. (2016). Automated detection of atrial fibrillation based on time–frequency analysis of seismocardiograms. *IEEE journal of biomedical and health informatics*, *21*, 1233–1241.

Kirley, K., Qato, D. M., Kornfield, R., Stafford, R. S., & Alexander, G. C. (2012). National trends in oral anticoagulant use in the united states, 2007 to 2011. *Circulation: Cardiovascular Quality and Outcomes*, *5*, 615–621.

Komorowski, M., & Celi, L. A. (2017). Will artificial intelligence contribute to overuse in healthcare? *Critical care medicine*, *45*, 912.

Kumar, D., Maharjan, R., Maxhuni, A., Dominguez, H., Frølich, A., & Bardram, J. E. (2021). mCardia: A Context-Aware Ambulatory ECG Collection System for Arrhythmia Screening. *ACM Transactions on Computing for Healthcare [Manuscript accepted for publication]*, .

Kwon, S., Hong, J., Choi, E.-K., Lee, E., Hostallero, D. E., Kang, W. J., Lee, B., Jeong, E.-R., Koo, B.-K., Oh, S. et al. (2019). Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study. *JMIR mHealth and uHealth*, *7*, e12770.

Lahdenoja, O., Hurnanen, T., Iftikhar, Z., Nieminen, S., Knuutila, T., Saraste, A., Kiviniemi, T., Vasankari, T., Airaksinen, J., Pänkäälä, M. et al. (2017). Atrial fibrillation detection via accelerometer and gyroscope of a smartphone. *IEEE Journal of Biomedical and Health Informatics*, *22*, 108–118.

Lai, D., Bu, Y., Su, Y., Zhang, X., & Ma, C.-S. (2020). Non-standardized patch-based ecg lead together with deep learning based algorithm for automatic screening of atrial fibrillation. *IEEE Journal of Biomedical and Health Informatics*, *24*, 1569–1578.

Lai, D., Zhang, X., Zhang, Y., & Heyat, M. B. B. (2019). Convolutional neural network based detection of atrial fibrillation combing rr intervals and f-wave frequency spectrum. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4897–4900). IEEE.

Limam, M., & Precioso, F. (2017). Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network. In *2017 Computing in Cardiology (CinC)* (pp. 1–4). IEEE.

Melgani, F., & Bazi, Y. (2008). Classification of electrocardiogram signals with support vector machines and particle swarm optimization. *IEEE transactions on information technology in biomedicine*, *12*, 667–677.

Minami, K.-i., Nakajima, H., & Toyoshima, T. (1999). Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network. *IEEE transactions on Biomedical Engineering*, *46*, 179–185.

Moody, G. (1983). A new method for detecting atrial fibrillation using rr intervals. *Computers in Cardiology*, (pp. 227–230).

Moody, G. B., & Mark, R. G. (2001). The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, *20*, 45–50.

Movisens GmbH (2019). EcgMove 4 – ECG and Activity Sensor. URL: https://www.movisens.com/en/products/ecg-sensor/.

Movisens GmbH (2020). DataAnalyzer-Sensor Data Analysis. URL: https://www.movisens.com/en/products/dataanalyzer/.

Nguyen, A., Ansari, S., Hooshmand, M., Lin, K., Ghanbari, H., Gryak, J., & Najarian, K. (2018). Comparative study on heart rate variability analysis for atrial fibrillation detection in short single-lead ecg recordings. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 526–529). IEEE.

Noh, Y.-H., Yap, J. H., & Jeong, D.-U. (2013). Implementation of the abnormal ecg monitoring system using heartbeat check map technique. In *2013 International Conference on IT Convergence and Security (ICITCS)* (pp. 1–4). IEEE.

Pan, J., & Tompkins, W. J. (1985). A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (pp. 230–236).

Parvaneh, S., Rubin, J., Babaeizadeh, S., & Xu-Wilson, M. (2019). Cardiac arrhythmia detection using deep learning: A review. *Journal of electrocardiology*, *57*, S70–S74.

Petmezas, G., Haris, K., Stefanopoulos, L., Kilintzis, V., Tzavelis, A., Rogers, J. A., Katsaggelos, A. K., & Maglaveras, N. (2021). Automated atrial fibrillation detection using a hybrid cnn-lstm network on imbalanced ecg datasets. *Biomedical Signal Processing and Control*, *63*, 102194. URL: http://www.sciencedirect.com/science/article/pii/S1746809420303323. doi:https://doi.org/10.1016/j.bspc.2020.102194.

Poh, M.-Z., Poh, Y. C., Chan, P.-H., Wong, C.-K., Pun, L., Leung, W. W.-C., Wong, Y.-F., Wong, M. M.-Y., Chu, D. W.-S., & Siu, C.-W. (2018). Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart*, *104*, 1921–1928.

Pourbabaee, B., Roshtkhari, M. J., & Khorasani, K. (2018). Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *48*, 2095–2104.

Press, W. H., & Teukolsky, S. A. (1990). Savitzky-golay smoothing filters. *Computers in Physics*, *4*, 669–672.

Sadr, N., Huvanandana, J., Nguyen, D. T., Kalra, C., McEwan, A., & de Chazal, P. (2016). Reducing false arrhythmia alarms in the icu using multimodal signals and robust qrs detection. *Physiological measurement*, *37*, 1340.

Sazonov, E. S., & Fontana, J. M. (2011). A sensor system for automatic detection of food intake through non-invasive monitoring of chewing. *IEEE sensors journal*, *12*, 1340–1348.

Shen, Y., Voisin, M., Aliamiri, A., Avati, A., Hannun, A., & Ng, A. (2019). Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1909–1916).

Shyu, L.-Y., Wu, Y.-H., & Hu, W. (2004). Using wavelet transform and fuzzy neural network for vpc detection from the holter ecg. *IEEE Transactions on Biomedical Engineering*, *51*, 1269–1273.

Sørensen, R., Gislason, G., Torp-Pedersen, C., Olesen, J. B., Fosbøl, E. L., Hvidtfeldt, M. W., Karasoy, D., Lamberts, M., Charlot, M., Køber, L. et al. (2013). Dabigatran use in danish atrial fibrillation patients in 2011: a nationwide study. *BMJ open*, *3*.

Surawicz, B., & Parikh, S. R. (2002). Prevalence of male and female patterns of early ventricular repolarization in the normal ecg of males and females from childhood to old age. *Journal of the American College of Cardiology*, *40*, 1870–1876.

Tran, L., Li, Y., Nocera, L., Shahabi, C., & Xiong, L. (2020). Multifusionnet: Atrial fibrillation detection with deep neural networks. *AMIA Summits on Translational Science Proceedings*, *2020*, 654.

Viskin, S., Golovner, M., Malov, N., Fish, R., Alroy, I., Vila, Y., Laniado, S., Kaplinsky, E., & Roth, A. (1999). Circadian variation of symptomatic paroxysmal atrial fibrillation. data from almost 10000 episodes. *European heart journal*, *20*, 1429–1434.

Wang, J. (2020). A deep learning approach for atrial fibrillation signals classification based on convolutional and modified elman neural network. *Future Generation Computer Systems*, *102*, 670–679.

Wei, X., Li, J., Zhang, C., Liu, M., Xiong, P., Yuan, X., Li, Y., Lin, F., & Liu, X. (2019). Atrial fibrillation detection by the combination of recurrence complex network and convolution neural network. *Journal of Probability and Statistics*, *2019*.

Wolf, P. A., Abbott, R. D., & Kannel, W. B. (1987). Atrial fibrillation: a major contributor to stroke in the elderly: the framingham study. *Archives of internal medicine*, *147*, 1561–1564.

Xia, Y., Wulan, N., Wang, K., & Zhang, H. (2018). Detecting atrial fibrillation by deep convolutional neural networks. *Computers in biology and medicine*, *93*, 84–92.

Xu, S. S., Mak, M.-W., & Cheung, C.-C. (2018). Towards end-to-end ecg classification with raw signal extraction and deep neural networks. *IEEE journal of biomedical and health informatics*, *23*, 1574–1584.

Yao, Z., Zhu, Z., & Chen, Y. (2017). Atrial fibrillation detection by multi-scale convolutional neural networks. In *2017 20th International Conference on Information Fusion (Fusion)* (pp. 1–6). IEEE.

Zhou, X., Ding, H., Ung, B., Pickwell-MacPherson, E., & Zhang, Y. (2014). Automatic online detection of atrial fibrillation based on symbolic dynamics and shannon entropy. *Biomedical engineering online*, *13*, 18.

Zoni-Berisso, M., Lercari, F., Carazza, T., & Domenicucci, S. (2014). Epidemiology of atrial fibrillation: European perspective. *Clinical epidemiology*, *6*, 213.

A.4 CACHET-CADB: A Contextualized Ambulatory Electrocardiography Arrhythmia Dataset

Authors
Devender Kumar, Sadasivan Puthusserypady, Helena Dominguez, Kamal Sharma, and Jakob E. Bardram

Journal
Scientific Data Nature. [Under Review]

CACHET-CADB: A Contextualized Ambulatory Electrocardiography Arrhythmia Dataset

Devender Kumar[1,*], Sadasivan Puthusserypady[1], Helena Dominguez[2], Kamal Sharma[3], and Jakob E. Bardram[1,*]

[1]Technical University of Denmark, Department of Health Technology, Kgs. Lyngby, Denmark
[2]Bispebjerg-Frederiksberg Hospital, Department of Cardiology, Copenhagen, Denmark
[3]U. N. Mehta Institute of Cardiology and Research Centre, Civil Hospital Campus, Ahmedabad, Gujarat, India
[*]Corresponding authors: Devender Kumar and Jakob E. Bardram ({deku,jakba}@dtu.dk)

ABSTRACT

Electrocardiogram (ECG) is a non-invasive tool for arrhythmia detection. In recent years, wearable ECG-based ambulatory arrhythmia monitoring has gained increasing attention. However, arrhythmia detection algorithms trained on existing public arrhythmia databases show higher afalse-positive rate (FPR) when applied to such ambulatory ECG recordings. It is primarily because the existing public databases are relatively clean as they are recorded using clinical-grade ECG devices in controlled clinical environments. They may not represent the signal quality and artifacts present in ambulatory patient-operated ECG. To help build and evaluate arrhythmia detection algorithms that can work on wearable ECG from free-living conditions, we present the design and development of the CACHET Contextualised Arrhythmia Database (CACHET-CADB), a multi-site contextualized ECG database from free-living conditions. In contrast to the existing databases, along with the ECG, CACHET-CADB also provides the continuous recording of patients' contextual data such as activities, body positions, movement accelerations, symptoms, stress level, and sleep quality. Currently, CACHET-CADB has 259 days of contextualized ECG recordings from 24 patients and 1602 manually annotated 10-seconds heart-rhythm samples.

Background and Summary

A heart arrhythmia like atrial fibrillation (AF) alone affects nearly 2% of the global adult population and is one of the major contributors to cardiovascular diseases (CVD) related morbid conditions and mortality[1,2]. The management of AF includes anti-coagulation to prevent strokes and heart rhythm-modifier medications[3]. However, for treatment to be effective in preventing further complications, early diagnosis and timely evaluation of AF plays a vital role. Analysis of electrocardiogram (ECG) signals is a non-invasive and cost-effective way of diagnosing AF. Due to their transient nature, paroxysmal AF remains under diagnosed in baseline ECGs and require long-term ECG monitoring. However, long-term preemptive monitoring is challenging as manual analysis of days/weeks-long ECG needed for detecting paroxysmal AF is resource and time-consuming.

Over the years, many computer-based algorithms have been developed for faster and accurate detection of AF and other types of arrhythmias[4]. More recently, with the advent of machine learning (ML) and deep learning (DL), the field of computer-aided AF analysis has experienced a huge breakthrough[4–6]. As compared to traditional ML and other feature engineering-based approaches, DL-based models can achieve end-to-end classification, thus removing the dependence on domain experts in the classification and stratification process. Despite all these advancements, one of the major challenge of using DL in AF classification is the availability of training and validation datasets. Although the DL algorithms can directly learn features from raw ECG data, it requires large and diverse datasets. The training data diversity helps the models to incorporate all the variations in inter/intra-personal ECG morphologies.

To meet this demand, many Internet ECG datasets such as the MIT-BIH AF Database (AFDB)[7], MIT-BIH Arrhythmia Database (MITDB)[8], PTB-LX[9], Computing in Cardiology Challenge 2017 Dataset (CinCDB)[10], Open-Access Arrhythmia Database (OA-ADB)[11], and DeepQ[12] have been published. Table 1 provides a summary of these publicly available arrhythmia databases. MITDB and AFDB are the earliest available ones and have been used extensively as a benchmark in training and evaluating ML/DL-based arrhythmia detection models[4,5,13].

Although the aforementioned databases have made a significant contribution for developing and evaluating arrhythmia detection models; generalisation and comprehensive performance evaluation of such models under free-living conditions remain questionable and face a number of significant challenges[4,13,14]:

Firstly, as mobile and wearable technology is advancing, wearable ECG devices have become available for longitudinal arrhythmia screening under free-living conditions. However, the majority of the current databases are either collected in

Table 1. Technical specifications and ECG annotation statistics of publicly available ECG databases. **Freq**: Sampling frequency (Hz); **Ch**: No. of ECG channels.

Database	Ch	Freq (Hz)	No. samples	Sample length	Rhythm classes	No. subjects	Context	Remark
AFDB[7]	2	250	23	10h	4	25	✗	continuous, controlled environment
MITDB[8]	2	360	48	30min	15	47	✗	continuous, controlled environment
NSRDB[15]	2	128	18	24h	1	18	✗	continuous, ambulatory
DeepQ[12]	1	250	897	5min	8	299	✗	intermittent, controlled environment
OA-ADB[11]	6	400	2000	30s	15	200	✗	continuous, ambulatory, patient-operated
CinC2017[10]	1	300	8528	9s to 60s	4	-	✗	intermittent, patient-operated
CACHET-CADB	1	1024	1602	10s	4	24	✓	continuous, ambulatory, patient-operated

controlled in-hospital settings or, in some cases, under the environments where patients are sitting without any motion. Therefore, the recordings are relatively clean and lack the ECG morphology changes and confounding artifacts that occur under free-living conditions. When the classification models trained on these datasets are applied to ambulatory wearable-based ECG recordings, they result in non-trivial false positives due to the degradation in the signal quality[16].

Secondly, the patient's context, such as physical activity and posture change, food intake (drinks or heavy meal), or mental stress, are known to introduce morphological changes in the ECG signal[17,18]. Existing databases only provide the raw ECG data, while information on the patient's context during the recording is missing. Recent systematic literature reviews of computer-aided arrhythmia analysis highlight that the arrhythmia detection in an ambulatory setting remains challenging and prone to mis-classification, without understanding the patient's context in which the ECG was undertaken[4,19]. Even during a manual ECG analysis, whenever a cardiologist finds 10 seconds or 30 seconds of ECG segment inconclusive, they often look for the longer context of the patient's ECG and rely on their knowledge about arrhythmia epidemiology[20]. Therefore, the patient's ambulatory context is essential for avoiding inappropriate classification due to "arrhythmia mimicking artifacts". Recent databases like DeepQ Arrhythmia Database (DeepQ)[12] have tried to address this problem by providing ECG recordings under the following three activity classes viz. sitting, walking, and lying down. These are, however, still a very limited set of activities and are recorded under circumstances that are very discordant from the real-world free-living ambulatory settings.

Thirdly, databases are usually generated from a single centre for a short time period (minutes or hours) on a homogeneous group of participants. Due to large variations that exist in the morphologies of ECG waveforms and the lack of diversity in current datasets, models trained on such datasets result in a large number of false positives when applied to ECG from different user contexts, ethnic characteristics, anthropomorphic features, gender, age group, and time-periods[4,21,22]. For instance, a multi-scale convolutional neural networks[21] showed a 98.18% accuracy when trained and validated on the AFDB, but its accuracy was reduced to 94.93% when applied on a Chinese dataset collected under free-living conditions. Similarly, the model by Andersen et al.[23] trained on AFDB has an excellent performance in 5-fold cross-validation on AFDB; however, it resulted in 4.9% FPR on previously unseen normal sinus rhythm (NSR) database from healthy individuals.

To complement the existing databases and to address some of the above-mentioned challenges, we present the CACHET-CADB. In contrast to the existing databases, CACHET-CADB provides the following unique features:

- It contains longitudinal wearable based ECG data from arrhythmia patients collected under *free-living conditions*, thus suitable for training and evaluating algorithms aimed at enabling real-time ambulatory ECG monitoring of the patients.
- Along with the ECG dataset, it also provides *contextual data* such as activities, body positions, movement accelerations, patient-reported events like symptoms experienced, sleep quality, stress level, and food intake. This contextualized ECG data can help make the end-to-end DL-based ECG classification models more explainable. Further, identifying the algorithm's source of errors in relation to the patient's ambulatory context can help in dynamically fine-tune it for those false-positives prone/inducing contexts under free-living conditions.
- Is multi-site and diverse (currently, Denmark and India but will be expanded further).

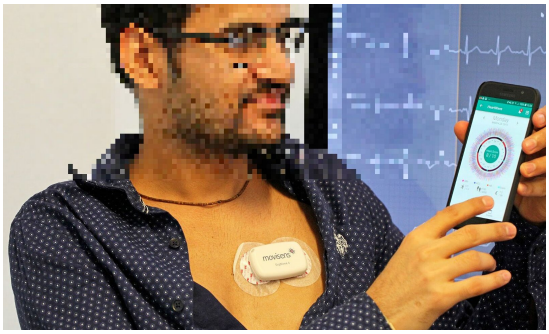


Figure 1. Data collection setup: (a) a chest-mounted single channel wireless ECG monitor collecting ECG and inertial (movement) measurements, and (ii) the mCardia mobile application for collection of patient-reported data[26].

Currently, the CACHET-CADB contains 259 days long contextualized ECG data from 24 patients. It also comprises 1602 annotations of 10-seconds long ECG-waveform, manually annotated by two independent Qualified cardiologists into four different heart rhythm classes: AF, NSR, 'noise', and 'other'. The CACHET-CADB is under continuous development, and annotations by cardiologists will be added to the database as they become available. The ECG annotation tool will be made public to increase the effort of crowd-sourcing the annotation process. Along with the dataset, a set of Python scripts and other software tools for data access, visualization, and data processing are available on the CACHET GitHub repository[24]. The dataset is freely available at DTU Data[25] at the Technical University of Denmark (DTU).

Methods

This section explains the data acquisition process, including ethical considerations, the data collection methods and technology, the data specifications, and the annotation process.

Data acquisition

Ethical consideration

The data for the CACHET-CADB was collected in India and Denmark. In Denmark, the study was exempted for ethical approval by the Danish Research Ethical Committee because the ECG recordings were only collected for technical purposes, and not to be used in a clinical setting (File # H-19071015). In India, the data collection was done with Mahatma Gandhi University of Medical Sciences & Technology (MGUMST), Jaipur, and the process complies with MGUMST's human participant's guideline and regulation as stated by the MGUMST Institutional Review Board (IRB). The approvals were granted on the ground that data collection was purely for technology development, and that the data would not be used for clinical diagnosis or treatment of the patients.

Recruitment

The participants were recruited during their out-patient arrhythmia clinic visits via a general announcement to participate in the data collection study. It was also made clear to participants that their participation was purely for research purposes, and the collected data would not be used in their ongoing clinical diagnosis or treatment. Preference was given to the participants with a known history of paroxysmal AF or high AF risk factors. All participants signed an informed consent form and allowed their data to be used and shared publicly after subject identity anonymization.

Data Collection Method

We used the mCardia system[26,27] for the data collection. It uses a single-channel chest-mounted wireless ECG Holter (the Movisens ECGMove4[28]) and a mobile application for data collection (Figure 1). Participants wore the ECG device using two disposable adhesive wet Ag/AgCl electrodes. All data was forwarded to, and stored in the CACHET Research Platform (CARP)[29], which is a secure and scalable cloud-based infrastructure for health data science hosted at DTU. Each participant installed the mCardia mobile application on his/her phone and continuously wore the ECG device for a minimum of 24 hours and up to 3 weeks. Participants were instructed to change the ECG electrodes daily and fill in the patient-reported information (symptoms, stress levels, sleep quality, and food intake) in the mCardia app. Further details on the mCardia system and CARP can be found at `https://carp.cachet.dk/mcardia/`.

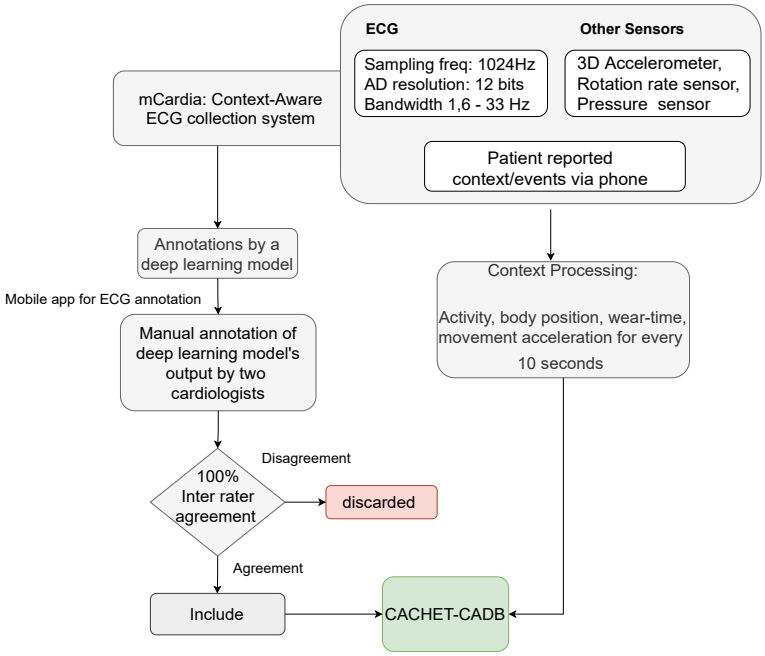


Figure 2. Overview of data collection and annotation process.

Anonymization and data trimming

The initial recording length varied from 24 hours to 3 weeks. For better manageability, analysis, and data handling, recordings were trimmed and assigned an anonymous ID (see Figure 5). In each record, the first (0th) and the last days are of variable lengths, whereas the rest are 24 hours long, starting from midnight.

ECG annotation process

Figure 2 shows the process used for annotating the ECG samples in the CACHET-CADB. A DL based AF detection model[23] was used to process the raw ECG recording. The AF onset and offset timestamps marked by the DL model were stored in CSV files. Thereafter, the segments between the onset and the offset were chopped into 10 seconds interval recordings and sent to two independent cardiologists for manual annotation via a mobile ECG annotation app. Figure 3 shows the user interface of the ECG annotation app used for the manual annotation. The annotation rules were discussed and agreed upon by the two cardiologists. A 10-second segment was assigned a label if it contained more than 50% of a particular rhythm type. If there were multiple rhythm classes in 10 seconds sample without having a majority ($\geq$50%) of a particular class, then it was annotated as "others". If artifacts in the 10-second signal precluded proper interpretation of the underlying rhythm, then the sample was annotated as "noise". The annotations of the two independent cardiologists were compared for inter-observer agreement. If there were disagreement between the two cardiologists, the annotations were discarded. Thus, the final database only includes samples where there is an agreement between the two cardiologist's annotations.

Processing Contextual data

The collected contextual data is of two categories: (1) patient-reported data collected via the mCardia app, and (2) sensor-generated data which is passively collected from the sensors on the mobile phone and the ECG Movisens device. Table 2 provides an overview of the types of collected data.

Patient-reported data

Patient-reported contextual data was collected when the patient manually enters data during the study period. We collected two types of patient-reported context information; (1) experienced events, and (2) daily health reports. The events were registered by patients when they experienced any unusual symptoms (e.g., palpitations, heartburn, etc.) during the ECG recording period.

Figure 3. Mobile application used for ECG annotation

Table 2. Specifications of the collected data. S: Sensed; PR: Patient-reported; EB: Event-based.

Collected data type	Type	Data source	Sampling rate
ECG	S	EcgMove4	1024 Hz
3D acceleration	S	EcgMove4	64Hz
Rotation rate sensor	S	EcgMove4	64Hz
Pressure sensor	S	EcgMove4	8Hz
Events	PR	EcgMove4 & Phone	EB
Sleep	PR & S	Phone	1/Day
Dietary	PR	Phone	1/Day

It includes details about the type of symptom, its duration, activity during the symptom, and a short note providing more context and experience. Health reports were provided daily and comprised of a three short survey on meals (timings and type of meal (light, heavy, moderate), self-perceived stress level, and sleep quality (on a scale of 1 to 5). The freestyle comments added by patients for further describing the symptoms or events were either in English or in the local vernacular language.

Sensor-generated data

The sensed context is passively derived from the on-board sensors (3D acceleration sensor, gyroscope, and pressure sensor) of the chest-mounted Movisens ECG device and from the phone's sensors. Table 2 lists the sensors' sampling rates. The DataAnalyzer Tool[30] was used for processing data from the Movisens sensors, and context data such as movement acceleration, body position, activity, step count, wear time, energy expenditure, and Metabolic Equivalent for Task (MET) levels were derived for an interval of 10 seconds. The movement acceleration, also known as Movement Acceleration Intensity (MAI), is a typical physical activity metric that depicts bodily movements' intensity. The MAI is measured in 'g', which is multiples of Earth's gravity (1g = 9.81 m/s2). In the DataAnalyzer Tool, the body positions were classified based on the inclination obtained from the 3D accelerometer. Its activity recognition is based on a white-box decision tree on the features extracted from the accelerometer and the barometric air pressure data[31]. The type of recognized activities include unknown, lying, sitting, standing, cycling, slope up, jogging, slope down, walking, and not-worn. Similarly, the body positions are classified based on the inclination obtained from the 3D accelerometer. The body position classes include unknown, lying supine, lying left, lying prone, lying right, upright, sitting/lying, standing, and not-worn.

Data Records

The CACHET-CADB includes over 259 days of single-channel contextualized ECG recording from 24 patients previously diagnosed with or suspected of the high risk of AF. Besides the patient's ambulatory contexts, it also contains 1602, 10-seconds long annotation samples of 4 different ECG rhythm classes, namely, AF, NSR, noise, and others (anything excluding AF, NSR, and noise). A sample of each of these rhythm classes is shown in Figure 4. The CACHET-CADB is freely available on DTU Data figshare[25] under the name "CACHET-CADB".

Figure 5 describes the organization of the records in CACHET-CADB. For better manageability and incorporation of future updates, the dataset is split into two main parts: (i) the raw signals (i.e., ECG, 3D accelerometer, angular rate) and (ii) the

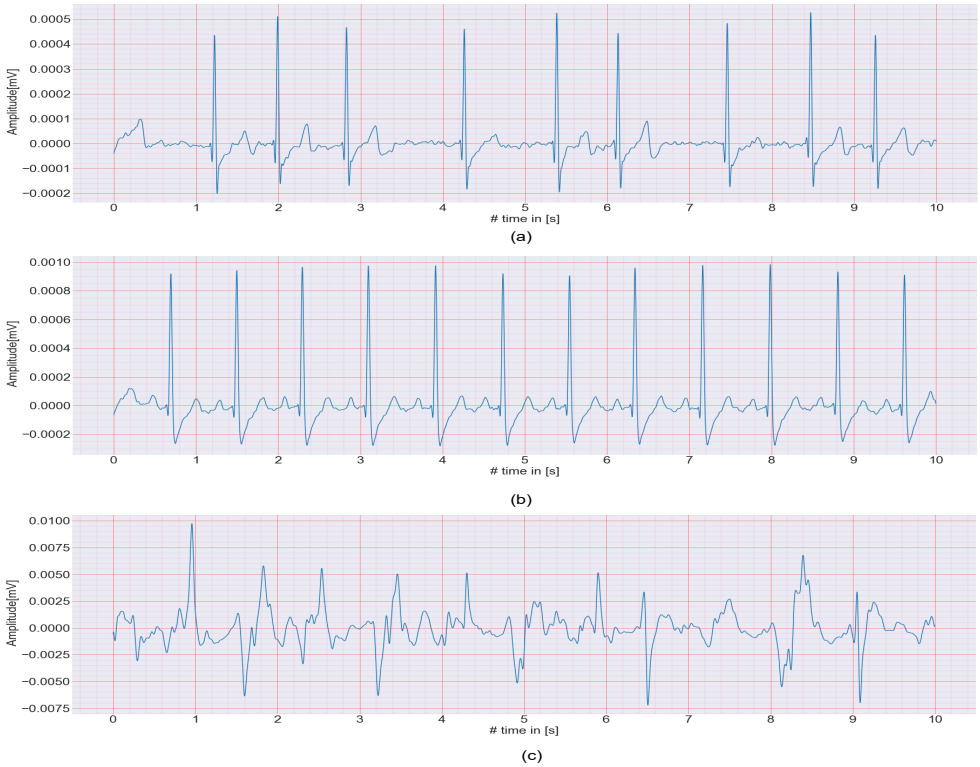


Figure 4. Figure (a), (b), and (c) show the 10 seconds ECG recordings of AF, NSR, and Noise classes, respectively.

annotations, while keeping the same folder structure inside each part. At the time of drafting this manuscript, the dataset has 24 records, spanning 259 days of recording from 24 patients of which, 7 were Danish and 17 were Indian. There were 15 males / 9 females – with an average age of 59, and of which 11 patients had documented one or more AF episodes in past.

Raw Signals and Metadata

The raw sensor data is stored in Unisens[32] file format. It allows simultaneously multi-sensor data, with synchronous storage at different sample rates, and comes with a human-readable meta-file in XML format. As illustrated in Figure 5, for each day the *unisens.xml* file contains the metadata for the raw signals. Table 3 describes these metadata in detail. The general metadata information includes the start timestamp, the total recording time (in seconds), and the anonymous user id (same as the anonymous id for the entire recording). The patient metadata includes height, weight, gender, location of the ECG sensor, and age at the time of recording. The raw ECG, 3D accelerometer, angular rate, and pressure signals are in the *ecg.bin*, *acc.bin*, *angularrate.bin*, and *press.bin* files, respectively. To allow for any future processing and analysis of the recordings, the dataset contains the raw signal without any preprocessing or filtering. However, given the recordings' ambulatory nature, any use of the data would probably need to implement baseline correction and removal of other artifacts beyond the normal ECG band [0.5–50Hz].

Annotations and Metadata

As shown in Figure 5, the annotations follow the same folder structures as the raw signals. For each day, the *context.xlsx* and *annotation.csv* files contain the contextual and annotation data, respectively. The *context.xlsx* file contains the patient's ambulatory context for every 10 seconds interval. These contextual data are derived from a 3D acceleration sensor, gyroscope, and pressure sensor, as described earlier. Table 4 provides the metadata for these contextual data, where the attributes listed in the table are columns in the *context.xlsx* file. The 'unit' column in Table 4 represents the measurement unit of each attribute. The remark column provides the label of each subclass within the same column. For instance, *ActicityClass* has several sub-classes,

Table 3. Metadata for the signal files described in the *unisens.xml* file of each record.

Type	Key	Data Type	Channel Name	Description
General	duration	string		Total recording time in seconds
	timestampStart	string		Recording start time
	measurementId	string		Anonymous user id
Patient & Device	height	string		Height in centimeters
	weight	string		Weight in kilograms
	sensorVersion	string		Recording device version
	sensorType	string		Recording device type
	age	string		age at recording in years
	sensorLocation	string		ECG sensors location on body
	personId	string		Anonymous user id
	gender	string		Gender (M/F)
ECG	ecg.bin	binary	ECG I	Resolution: 12 bit, Input range CM = 560 mV, DM = +/-5 mV, 3db bandwidth 1,6 - 33 Hz Output rate: 1024 Hz
Accelerometer	acc.bin	binary	accX, accY, accZ	3D acceleration sensor Measurement range: +/- 16 g Output rate: 64 Hz
Angular Rate	angularrate.bin	binary	angularRateX, angularRateY, angularRateZ	Rotation rate sensor: Measurement range: +/-2000 dps Output rate: 64 Hz
Pressure	press.bin	binary	press	Measurement range: 300 - 1100 hPa Noise: 0,03 hPa Output rate: 8 Hz
Marker	marker.csv	integer		Contains indexes of *events* when the patient experienced unusual systems and tapped on ECG Holter. Divide the index by 64 to get the event time in seconds from the start of the recording. Output rate: 64 Hz

Table 4. Contextual-data descriptor table. The attributes are the columns of the *context.xlsx* file in the annotation folder of each day.

Attribute	Unit	Remark
Time rel	[s]	Relative time from start of measurements in seconds
Day rel	[d]	Number of days from start of measurement
Time rel	[hh:mm:ss]	Relative time from start of measurement
Date abs	[yyyy-mm-dd]	Absolute date
Time abs	[hh:mm:ss]	Absolute time
ActivityClass	-	Activity Class (0=unknown, 1=lying, 2=sitting/standing, 3=cycling, 4=slope up, 5=jogging, 6=slope down, 7=walking, 8=sitting/lying, 9=standing, 10=sitting/lying/standing, 11=sitting, 99=not worn)
ActivityEnergyExpenditure	[kcal/d]	Activity energy expenditure (AEE) in kcal/d
Altitude	[m]	Altitude from barometer
BodyPosition	-	Body position (0=unknown, 1=lying supine, 2=lying left, 3=lying prone, 4=lying right, 5=upright, 6=sitting/lying, 7=standing , 99=not worn)
InclinationDown	[deg]	Inclination of sensor axis down against the vertical (0 to 180 deg)
InclinationForward	[deg]	Inclination of sensor axis forward against the vertical (0 to 180 deg)
InclinationRight	[deg]	Inclination of sensor axis right against the vertical (0 to 180 deg)
MET		MET value directly calculated from regression models
MovementAcceleration	[g]	MovementAcceleration: Raw acceleration, bandpass filtered, vector magnitude
NonWearSleepWake	-	Sleep/Wake detection (0=wake, 1=sleep, 2=not worn)
NonWearTime	-	Non wear detection (0=worn, 1=not worn)
StepCount	[steps]	Count of steps per output interval
TotalEnergyExpenditure	[kcal/d]	Total energy expenditure (TEE = BMR + AEE)
VerticalSpeed	[m/s]	Vertical speed, calculated from barometer

Table 5. ECG annotation overview showing the class of rhythm types , its code in the *annotation.csv* file, and the number of available annotations for each class.

Class	Code	#
AF	1	747
NSR	2	615
Noise	3	221
Others	4	19

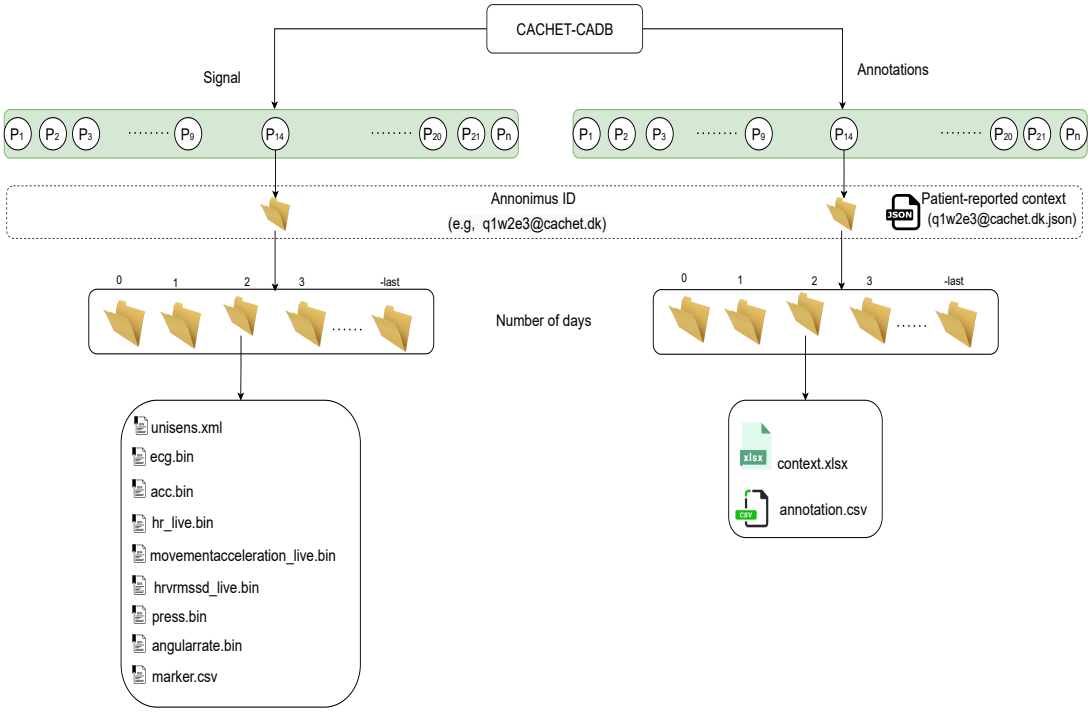


Figure 5. The structure of the data records in CACHET-CADB. Overall, the database is divided into two major parts; (i) the raw recordings in binary files and (ii) the contextual information including patient-reported data and the annotations. Each record is organized according to patient ID first and day in study subsequently.

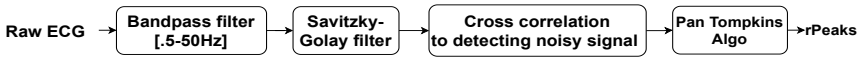


Figure 6. Analysis of ECG quality, QRS complex and R-peak detection.

such as lying, sitting/standing, cycling, slope up, or jogging. The corresponding subclass code (0, 1, 2...) represents them in the activity column of the *context.xlsx* file. Patient-reported data is provided as a single JSON file in each annotation folder (see Figure 5). The JSON file contains two types of data *"dailyInfo"*, and *"event"*. Their metadata are described in Table 6 (a) and Table 6 (b), respectively.

The *annotaion.csv* file contains the cardiologists' annotation of hearth rhythms. It contains the following columns: (i) the start index of 10 seconds long segment (*Start*), (ii) the end index of 10 seconds long segment (*End*), and (iii) the ECG rhythm class (*Class*). Table 5 provides the statistics of each of the annotated rhythm classes and their associated code in the *Class* column of the *annotation.csv* file.

Technical Validation

Quality assessment of ECG annotation

Although the DL models[23] was used for automatic labeling (Figure 3), to ensure the quality and integrity of the rhythm annotation, we have released only the annotations that have been manually checked by the two independent cardiologists. A 100% inter-rater agreement policy is followed. The ECG segments on which there was a disagreement between two cardiologists are not included in this release.

Table 6. Metadata of patient-entered context data

(a) Metadata of *"dailyInfo"* field in JSON file.

Field name	Description
date_time	Day for which the *"dailyInfo"* is filled
bed_time	Bed time
awake_time	Wake up time
sleep_quality	Self-assessed sleep quality (1–5)
stress_level	Self-assessed stress level (1–5)
lunch_time	Lunch time
lunch_weight	Lunch quantity (heavy, moderate, light)
breakfast_time	Breakfast time
breakfast_weight	Breakfast quantity (heavy, moderate, light)
dinner_time	dinner time
dinner_weight	Dinner quantity (heavy, moderate, light)
other_time	Time of any other meal/drink
other_weight	Meal/Drink quantity (heavy, moderate, light)

(b) Metadata of *"event"* field in JSON file representing patient-reported symptoms that the patient may have experienced during the recording period.

Field name	Description
id	Unique id
notes	Note describing the unusual experience/symptoms
labels	n/a
source	How was the event entered? *"Tap"*: By tapping on the ECG Holter *"Self input"*: Manually created in the app
deleted[a]	Was the event deleted? (true/false)
comments	n/a[b]
duration	Time in seconds for which symptoms lasted
symptom	Symptom experienced during the unusual event (e.g., "Dizziness")
activity	Patients activity when the unusual symptoms were experienced
completed	Were the details of an event filled in? True: All fields were completed. False: Not filled/ Partially filled
reviewed	n/a
date_time	Time of the event as experienced by the patient

[a]The patient could delete an event e.g., if it was created by accidentally tapping the ECG device.
[b]The patient's comments are removed for anonymity.

Signal quality assessment

For testing the validity of the collected ECG data, an ECG signal quality assessment was done using an auto-correlation-based noise detector. Subsequently, the Pan Tomkinson algorithm[33] was used to calculate QRS complex/R-peaks. The steps used in the validation process are shown in Figure 6. As the ambulatory ECG signal tends to get contaminated by noise and other artifacts, first, a band-pass [0.5–50Hz] filter was applied, and the baseline was removed. A Savitzky–Golay[34] filter followed this to smoothen out the data. Thereafter, the signal was chopped into 10 seconds long windows, and an auto-correlation based noise detector was applied to detect the noisy signal. Finally, the Pan Tomkinson algorithm[33] was used to calculate the QRS complexes and the R-peaks for each of these 10 seconds windows. Table 7 shows the number of R-peaks detected and the percentage of the noisy signal detected in each record. In the ECG signal, intervals between the R-peak indicate heart rhythm's regularity. These RR intervals (RRI) features have been extensively used in DL-based AF detection models[23].

Although we did identify noise in the dataset, we did not exclude the noise from the database. This was done intentionally to allow the CACHET-CADB to reflect a realistic distribution of ECG quality as expected under free-living conditions. ECG riddled with confounding artifacts and varying signal quality is expected when performing longitudinal ambulatory arrhythmia screening. Therefore, we put forward the CACHET-DB as a resource for designing and evaluating DL-based arrhythmia detection algorithms, which work under free-living condition without generating false positives. Moreover, the database can be used for creating unsupervised learning methods, which can enable feature extraction representing ECG quality variation in ambulatory settings. As already discussed, one of the main challenges with the existing arrhythmia ECG datasets is that they are collected in a clinically controlled environment and are relatively clean. Models trained on such clean datasets may result in many false-positive cases when applied on ECG collected under free-living conditions that inevitably has low signal quality and many artifacts[35,36].

Table 7. Signal quality assessments and detection of QRS complex/R-peaks

Record	User Id	Days	No of R-peaks	Signal Duration (hours)	Noisy Signal (%)	Non-wear Time (hours)
P1	a2b3c4@cachet.dk	12	1158069	241.58	7.48	6.15
P2	t1y2u3@cachet.dk	7	673950	139.40	6.47	1.03
P3	q1w2e3@cachet.dk	15	1440323	315.77	8.40	41.08
P4	p1q2w3@cachet.dk	8	739199	173.14	5.80	10.85
P5	b1t2s3@cachet.dk	8	665666	147.97	16.27	25.50
P7	k9v3r7@cachet.dk	12	913892	260.16	12.43	41.91
P6	s1a2n3@cachet.dk	12	1241040	257.34	3.26	8.82
P8	g4v3r7@cachet.dk	22	2895927	479.16	9.90	77.98
P9	c1x2p3@cachet.dk	12	921713	247.78	29.61	82.21
P10	k1x2p3@cachet.dk	16	1297163	359.85	31.72	80.28
P11	v2c3r4@cachet.dk	16	1363671	326.96	12.72	61.26
P12	r4p2n8@cachet.dk	14	1988086	308.91	6.88	6.31
P13	f7c4n6@cachet.dk	19	1964554	412.19	2.65	16.63
P14	j4y9x6@cachet.dk	12	1035832	262.94	29.90	111.36
P15	u3h6c1@cachet.dk	14	1385906	315.49	28.05	79.08
P16	i6t2v4@cachet.dk	17	1567938	359.86	6.29	25.71
P17	z2y4b9@cachet.dk	15	1280062	325.34	6.02	19.18
P18	g2v5x7@cachet.dk	5	431256	92.95	3.23	1.54
P19	m1t2a3@cachet.dk	4	272549	75.22	3.59	2.51
P21	y1t2r3@cachet.dk	8	778148	168.93	10.34	12.10
P23	m1n2b3@cachet.dk	7	762802	160.54	7.24	6.33
PNSR-1	deku_test@cachet.dk	1	105079	24.00	0.49	0.56
PNSR-3	j5f3c2@cachet.dk	1	92134	26.44	27.14	0.00
PNSR-4	w1y3n2@cachet.dk	2	191867	48.00	5.63	2.05
Total		259	25166826	5529.94		726.57

Discussion

This paper presents the design and development of a contextualised ECG database to support the development and generalisation of ECG analysis and arrhythmia detection models. The CACHET-CADB has been developed as a part of the REAFEL[37] research project, which focuses on building mHealth and DL-based solutions for optimizing diagnosis of AF in the frail and elderly population. CACHET-CADB is particularly important for researchers who are working on bringing ECG analysis and AF detection on patient-operated wearable ECG into widespread adoption under free-living conditions. The database will be further expanded with more recordings and ECG annotation as they become available by following the data annotation and storage setup described above.

The ability to bring arrhythmia detection models in widespread adoption under free-living conditions is limited by the lack of a patient-operated ambulatory ECG dataset that truly represents all the confounding contamination expected in such conditions. The models trained on benchmark datasets in Table 1 show high performance when tested on the same datasets or similar datasets collected under clinical supervision. However, the high classification performances often obtained on these datasets are not reproducible when applied to patient-operated ECG data under free-living conditions. The patients-operated wearable-based ECG under free-living condition is often contaminated with arrhythmia mimicking artifacts and suffers from poor signal quality. The cause of the poor performance under free-living conditions is attributed to the lack of diversity and relatively good signal quality of ECG wave forms in these benchmark databases[16].

With wearable technology advancements, single lead portable ECG monitoring has been gained attraction for arrhythmia screening under free-living conditions[38]. Coupling portable patient-operated ECG monitoring with computer-aided ML and DL-based classification algorithms can help in real-time and cost-effective longitudinal arrhythmia screening under free-living conditions. To achieve high sensitivity and reproducibility under free-living conditions, the CACHET-CADB provides an opportunity to train and evaluate arrhythmia detection models on a dataset representing all the ECG morphology changes and confounding noise contamination expected in free-living conditions.

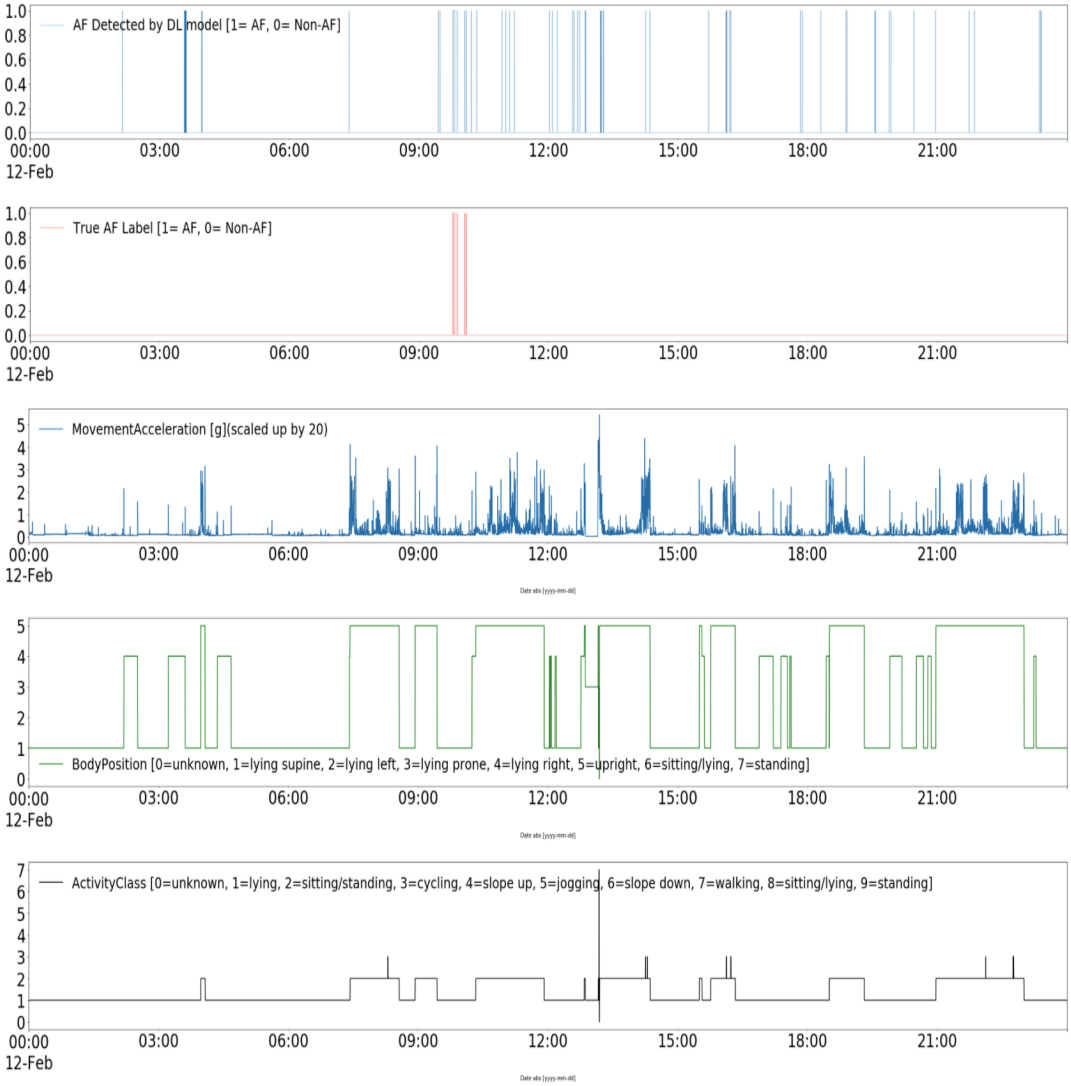


Figure 7. Explainable deep learning: This Figure shows a contextualized view of a deep learning-based AF detection model's performance on a single day of ECG from CACHET-CADB. In 24 hours of ECG under free-living conditions, short segments of false positive in a model's output are linked to change in activity, change in body position, and sudden movement accelerations.

Context-aware ECG for explainable DL models

One advantage of CACHET-CADB over the existing database is the availability of patients' ambulatory context corresponding to the recorded ECG. In the absence of patients' context, the ECG analysis under free-living conditions is prone to misclassification and misinterpretation[4]. The contextual data can also be used for multi-model input and context-based heuristics to dynamically fine-tune the models' sensitivity and specificity under different user contexts in ambulatory settings. To reduce the FPR, algorithms should be made adaptive to the user's context–i.e., the sensitivity and specificity of algorithms should be dynamically adjustable. For instance, in the elderly population, there is a significantly higher prevalence of falls in patients with AF[39]. Suppose an algorithm is applied to elderly patients' data and if a fall is detected, then the algorithm should factor-in for the fall in the dynamic adjustment of its sensitivity and specificity. Similarly, information about AF triggering contexts[40] such as high stress-level, food-intake (heavy meal), drinks (alcohol, caffeine) can be utilized to make algorithms more sensitive in those contexts.

Furthermore, the contextual data can pave the way for improving the interpretability of ML and DL models[41]. For instance, Figure 7 shows a DL model's AF classification results, the 'ground truth' annotations, and patient's ambulatory contexts (body position, activities, movement acceleration) for 24 hours long record in CACHET-CADB. It can be inferred from Figure 7 that the model is resulting in more false positives (FP) whenever there is a change in activity, body position, and movement acceleration, which is most prominent after 09:00 o'clock. Such information can be made available to a cardiologist for the manual inspection of the dataset thereby providing a better insight into when and why the AF detection algorithm has identified an AF episode. The information can also be utilized to build post-processing heuristics around these FP prone ambulatory contexts. With CACHET-CADB, we aim to provide the DL research community rich longitudinal contextualized ECG data that can help build and evaluate models which realistically work on patient-operated ECG from free-living ambulatory conditions.

Usage Notes

The design, data-descriptor, and the software tools for using CACHET-CADB are presented and made available for public use. When using this database, please cite the current publication. The new data recording and ECG annotations on the existing records will be added to CACHET-CADB periodically when they become available; details of the subsequent release will be available at CACHET's website[42].

Code availability

Visual inspection and editing of records can be done using the UnisensViewer tool `http://software.unisens.org/download/UnisensViewer/UnisensViewer_Setup.exe`. Python library *pyunisens* (`https://github.com/Unisens/pyunisens`) can be used for reading and editing the signal programmatically. We also provide a basic code example and Jupyter Notebook in Python for using the database `https://github.com/cph-cachet/cachet-ecg-db`. The contextual data file *context.xlsx* can be loaded and viewed using the panda library (`https://pandas.pydata.org/`); an example code for the same can be found at `https://github.com/cph-cachet/cachet-ecg-db`. All software is open sourced under an MIT license and we welcome pull requests.

References

1. Ball, J., Carrington, M. J., McMurray, J. J. & Stewart, S. Atrial fibrillation: profile and burden of an evolving epidemic in the 21st century. *Int. journal cardiology* **167**, 1807–1824, https://doi.org/10.1016/j.ijcard.2012.12.093 (2013).
2. Members, A. F. *et al.* 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: An update of the 2010 ESC Guidelines for the management of atrial fibrillationDeveloped with the special contribution of the European Heart Rhythm Association. *Eur. Hear. J.* **33**, 2719–2747, 10.1093/eurheartj/ehs253 (2012). https://academic.oup.com/eurheartj/article-pdf/33/21/2719/17893799/ehs253.pdf.
3. Fuster, V. *et al.* Acc/aha/esc 2006 guidelines for the management of patients with atrial fibrillation: A report of the american college of cardiology/american heart association task force on practice guidelines and the european society of cardiology committee for practice guidelines (writing committee to revise the 2001 guidelines for the management of patients with atrial fibrillation): Developed in collaboration with the european heart rhythm association and the heart rhythm society. *Circulation* **114**, e257–e354, https://doi.org/10.1161/CIRCULATIONAHA.106.177292 (2006).
4. Dinakarrao, S. M. P., Jantsch, A. & Shafique, M. Computer-aided arrhythmia diagnosis with bio-signal processing: A survey of trends and techniques. *ACM Comput. Surv. (CSUR)* **52**, 1–37, https://doi.org/10.1145/3297711 (2019).
5. Faust, O., Ciaccio, E. J. & Acharya, U. R. A review of atrial fibrillation detection methods as a service. *Int. J. Environ. Res. Public Heal.* **17**, 3093, https://doi.org/10.3390/ijerph17093093 (2020).

6. Matias, I. *et al.* Prediction of atrial fibrillation using artificial intelligence on electrocardiograms: A systematic review. *Comput. Sci. Rev.* **39**, 100334, https://doi.org/10.1016/j.cosrev.2020.100334.

7. Moody, G. A new method for detecting atrial fibrillation using rr intervals. *Comput. Cardiol.* 227–230, https://doi.org/10.13026/C2MW2D (1983).

8. Moody, G. B. & Mark, R. G. The impact of the mit-bih arrhythmia database. *IEEE Eng. Medicine Biol. Mag.* **20**, 45–50, https://doi.org/10.13026/C2F305 (2001).

9. Wagner, P. *et al.* Ptb-xl, a large publicly available electrocardiography dataset. *Sci. data* **7**, 1–15, https://doi.org/10.1038/s41597-020-0495-6 (2020).

10. Clifford, G. D. *et al.* Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, 1–4, 10.22489/CinC.2017.065-469 (IEEE, 2017).

11. Shen, Q. *et al.* An open-access arrhythmia database of wearable electrocardiogram. *J. Med. Biol. Eng.* **40**, 564–574, https://doi.org/10.1007/s40846-020-00554-3 (2020).

12. Wu, M.-H. & Chang, E. Y. Deepq arrhythmia database: A large-scale dataset for arrhythmia detector evaluation. In *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, 77–80, https://doi.org/10.1145/3132635.3132647 (2017).

13. Parvaneh, S., Rubin, J., Babaeizadeh, S. & Xu-Wilson, M. Cardiac arrhythmia detection using deep learning: A review. *J. electrocardiology* **57**, S70–S74, https://doi.org/10.1016/j.jelectrocard.2019.08.004 (2019).

14. Liu, F. *et al.* An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *J. Med. Imaging Heal. Informatics* **8**, 1368–1373, https://doi.org/10.1166/jmihi.2018.2442 (2018).

15. Goldberger, A. L. *et al.* Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* **101**, e215–e220, https://doi.org/10.13026/C2NK5R (2000).

16. Gao, H. *et al.* An open-access ecg database for algorithm evaluation of qrs detection and heart rate estimation. *J. Med. Imaging Heal. Informatics* **9**, 1853–1858, https://doi.org/10.1166/jmihi.2019.2800 (2019).

17. Agrafioti, F. & Hatzinakos, D. Ecg biometric analysis in cardiac irregularity conditions. *Signal, Image Video Process.* **3**, 329 (2009).

18. Van Dam, P., Mouton, S. & Oosterhoff, P. Template matching method for monitoring of ecg morphology changes (2011). US Patent 7,996,070.

19. Ebrahimi, Z., Loni, M., Daneshtalab, M. & Gharehbaghi, A. A review on deep learning methods for ecg arrhythmia classification. *Expert. Syst. with Appl. X* 100033, https://doi.org/10.1016/j.eswax.2020.100033 (2020).

20. Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. medicine* **25**, 65, https://doi.org/10.1038/s41591-018-0268-3 (2019).

21. Yao, Z., Zhu, Z. & Chen, Y. Atrial fibrillation detection by multi-scale convolutional neural networks. In *2017 20th International Conference on Information Fusion (Fusion)*, 1–6, 10.23919/ICIF.2017.8009782 (2017).

22. Ceylan, R. & Özbay, Y. Comparison of fcm, pca and wt techniques for classification ecg arrhythmias using artificial neural network. *Expert. Syst. with Appl.* **33**, 286–295 (2007).

23. Andersen, R. S., Peimankar, A. & Puthusserypady, S. A deep learning approach for real-time detection of atrial fibrillation. *Expert. Syst. with Appl.* **115**, 465–473, https://doi.org/10.1016/j.eswa.2018.08.011 (2019).

24. CACHET-CADB ToolKit. *Available [Online]:* https://github.com/cph-cachet/cachet-ecg-db (2021).

25. Kumar, D., Puthusserypady, S. & Bardram, J. E. CACHET-CADB. *Available [Online]:* https://doi.org/10.11583/DTU.14547264.v1 (2021).

26. Kumar, D. *et al.* mCardia: A Context-Aware Ambulatory ECG CollectionSystem for Arrhythmia Screening. *ACM Transactions on Comput. for Healthc.* ([Manuscript accepted for publication]).

27. Feasibility Study of mCardia: A Context-aware ECG monitoring system for arrhythmia screening. *Available [Online]:* https://www.cachet.dk/research/studies/mcardia (2019).

28. Movisens. EcgMove4 – ECG and Activity Sensor. *Available [Online]:* https://www.movisens.com/en/products/ecg-sensor/ (2020).

29. CACHET Research Platform (CARP). *Available [Online]:* https://carp.cachet.dk/ (2019).

30. Movisens. Dataanalyzer-sensor data analysis. https://www.movisens.com/en/products/dataanalyzer/ (2020).

31. Anastasopoulou, P., Tansella, M., Stumpp, J., Shammas, L. & Hey, S. Classification of human physical activity and energy expenditure estimation by accelerometry and barometry. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6451–6454, 10.1109/EMBC.2012.6347471 (2012).

32. Movisens. Unisens file format [online], https://docs.movisens.com/Unisens/UnisensFileFormat/ (2021).

33. Pan, J. & Tompkins, W. J. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering* 230–236, 10.1109/TBME.1985.325532 (1985).

34. Press, W. H. & Teukolsky, S. A. Savitzky-golay smoothing filters. *Comput. Phys.* **4**, 669–672 (1990).

35. Halvaei, H., Svennberg, E., Sörnmo, L. & Stridh, M. False alarm reduction in atrial fibrillation screening. In *2020 Computing in Cardiology*, 1–4, 10.22489/CinC.2020.255 (IEEE, 2020).

36. Fan, X. *et al.* Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ecg recordings. *IEEE journal biomedical health informatics* **22**, 1744–1753 (2018).

37. REAFEL: Reaching the Frail Elderly Patient for optimizing diagnosis of atrial fibrillation. *Available [Online]:* https://www.cachet.dk/research/research_projects/REAFEL (2020).

38. Ramkumar, S. *et al.* Atrial fibrillation detection using single lead portable electrocardiographic monitoring: a systematic review and meta-analysis. *BMJ open* **8**, e024178 (2018).

39. Hung, C.-Y. *et al.* Falls and atrial fibrillation in elderly patients. *Acta Cardiol. Sinica* **29**, 436 (2013).

40. Groh, C. A. *et al.* Patient-reported triggers of paroxysmal atrial fibrillation. *Hear. rhythm* **16**, 996–1002, https://doi.org/10.1016/j.hrthm.2019.01.027 (2019).

41. Meira Jr, W., Ribeiro, A. L., Oliveira, D. M. & Ribeiro, A. H. Contextualized interpretable machine learning for medical diagnosis. *Commun. ACM* **63**, 56–58, http://dx.doi.org/10.1145/3416965 (2020).

42. The CACHET Contextualized Arrhythmia Database (CACHET-CADB). *Available [Online]:* https://www.cachet.dk/resources/data/cachet-cadb (2021).

Acknowledgements

The authors wish to thank the team headed by Dr. Rajeev Sharma, Head of Electrophysiology at Heart failure division at Mahatma Gandhi University of Medical Sciences & Technology (MGUMST), Jaipur India, for their support. The authors gratefully thank Maj Gen Dr. A.K. Singh (Retd) and Prof Bipin kumar Rathod from the Department of Health Informatics at MGUMST for helping in the data collection process and patient recruitment. This research has been funded by the Innovation Fund Denmark as part of the REAFEL project (IFD Project No. #6153–00009B) and the Copenhagen Center for Health Technology.

Author contributions statement

DK and JEB conceived the database and implemented the technology for data collection and storage; DK conducted the data collection and manual ECG annotation process in collaboration with HD and KS. DK analyzed the data and wrote the Python scripts. DK, JEB, and SP wrote the paper. All authors reviewed the manuscript. JEB obtained the funding.

Competing interests

The authors declare no competing interests.

A.5 DeepAware: A Hybrid Deep Learning and Context-Aware Heuristics Based Model for Atrial Fibrillation Detection

`Authors`
Devender Kumar, Abdolrahman Peimankar, Sadasivan Puthusserypady, Helena Dominguez, Kamal Sharma, and Jakob E. Bardram

`Journal`
[Manuscript in preparation]

DeepAware: A Hybrid Deep Learning and Context-Aware Heuristics Based Model for Atrial Fibrillation Detection

Devender Kumar, Abdolrahman Peimankar, Kamal Sharma, Helena Dominguez, Sadasivan Puthusserypady, and Jakob E. Bardram

***Abstract*— The automatic atrial fibrillation (AF) detection models trained on RR-interval (RRI) based features usually achieve a high performance on the standard benchmark ECG datasets. However, they may result in significant false positives (FPs) when tested on patient-operated single-channel ECG data from free-living ambulatory conditions and in the presence of non-AF arrhythmias. In this paper, to alleviate such false positive rate (FPR) and to improve the AF detection performance on free-living ambulatory ECG, we propose *DeepAware* – a hybrid deep learning (DL) and context-aware heuristics-based model for AF detection. It combines the ventricular response (i.e., RRI) features, atrial activity (i.e., P-wave) features, and a context-aware-heuristics model to improve the AF detection and significantly reduces the FPR. The proposed model has outperformed the state-of-the-arts on the standard benchmark ECG datasets. To show the capability of the model, it has also been evaluated on a private dataset CACHET Contextualised Arrhythmia Database (CACHET-CADB) from free-living conditions, on which it achieved a sensitivity (Se), specificity (Sp), and accuracy (Acc) of 97.94%, 98.39%, 98.06%, respectively. Further, we compared *DeepAware*'s performance with the RRI features-based AF detection model and demonstrated that incorporating atrial activity features and context-awareness can significantly reduce the FPR caused by confounding non-AF arrhythmias and ambulatory context changes. By lowering the FPR, the *DeepAware* model can substantially reduce the physician's workload of manually reviewing the FPs.**

***Index Terms*— Atrial fibrillation, convolutional neural networks (CNNs), deep learning, electrocardiogram (ECG), health informatics, long short-term memory (LSTM)**

I. Introduction

Atrial fibrillation (AF) is one of the most prevalent types of cardiac arrhythmias. It is considered a leading cause of stroke and other heart-related complications in elderly population [1]. Nearly 2.3 million people in the USA alone are affected by AF, and this number is likely to increase by 2.5 times by the year 2050 [1]. Early diagnosis and anti-coagulation medication can help in preventing AF complications [2]. The electrocardiogram (ECG) analysis is one of the most inexpensive and non-invasive ways for AF detection. However, due to its abrupt and paroxysmal nature, it may be very challenging to detect AF during routine in-hospital ECG checkups. Thus, a longitudinal screening in a patient's natural setting is needed. One of the most common ways to analyze ECG signals by physicians and cardiologists is through visual examination of the recordings. However, it is a very difficult and time consuming task to analyze such huge amounts of data. Therefore, it is of great interest to develop reliable software to analyze and interpret ECG signals to detect cardiac arrhythmias.

Various state-of-the-art algorithms have been introduced in the literature to automatically detect AF from ECG recordings [3]. Most of these algorithms are based on classical machine learning and feature engineering techniques (e.g., temporal intervals, wavelet transform, etc.) [4]. Feature engineering is an essential step in these models to transform raw data into a suitable representation as inputs for the machine learning model to distinguish between different cardiac arrhythmias. Even though feature engineering based algorithms perform very well in some cases, they face three main challenges: (1) they require hand-crafted feature extraction by a domain expert, (2) they are susceptible to noise in ambulatory setting, and (3) they have relatively low generalization on new data [3]–[5].

In recent years, there have been many breakthroughs in various applications of DL in the areas such as computer vision, natural language processing, and health informatics [6]–[10]. In addition, DL has been widely explored to analyze ECG signals to detect AF in heart disease patients. Various end-to-end DL models have been introduced for AF detection, which basically bypass the handcrafted feature engineering step needed by other machine learning methods [5], [11]–[19]. Wang [14] proposed a convolutional neural network (CNN) and a modified Elman neural network (MENN) based AF detection model, which achieved an accuracy of 97.4% on MIT-BIH AF Database (AFDB) dataset. Similarly, Faust et al. [13] applied a LSTM model for AF detection on heart

This work was supported in part by the Innovation fund Denmark under grant # 6153-00009B (REAFEL).

D. Kumar, S. Puthusserypady and J. E. Bardram are with the Department of Health Technology, Technical University of Denmark, 2800 Lyngby, Denmark (e-mail: deku@dtu.dk, jakba@dtu.dk, sapu@dtu.dk).

A. Peimankar is with SDU Health Informatics and Technology, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, 5230 Odense, Denmark (e-mail: abpe@mmmi.sdu.dk).

[2]Helena Dominguez is with the Bispebjerg-Frederiksberg Hospital, Department of Cardiology, Copenhagen, Denmark (e-mail: maria.helena.dominguez.vall-lamora.02@regionh.dk).

[3]Kamal Sharma, is with the U. N. Mehta Institute of Cardiology and Research Centre, Civil Hospital Campus, Ahmedabad, Gujarat, India (e-mail: kamalcardiodoc@gmail.com).

rate features, which achieved 98.51% accuracy using 10-fold cross-validation on AFDB dataset. Petmezasa et al. [20] built a hybrid CNN-LSTM model that utilizes focal loss when dealing with an imbalance train set. This model also performs well on AFDB dataset with a high sensitivity and specificity of 97.87% and 99.29%, respectively.

Despite the promising performance of the above-mentioned research on the publicly available datasets, applying them for longitudinal AF screening under free-living conditions still remains an open challenge for several reasons [3], [21].

Firstly, most of the DL models have been built and evaluated on the public databases, which are relatively clean and contain manually corrected/annotated R-peaks [22], [23]. The AF detection algorithms based on RR-interval (RRI) features are limited by their assumption of receiving almost perfect R-peak detection [22]. On the other hand, due to the users' continuous movements in free-living conditions, the patient-operated ECG recordings can often be confounded with various AF mimicking artifacts and noises [24]. Presence of such noises makes the detection of R-peaks and P-waves cumbersome if not impossible in some cases. Consequently, this result in non-trivial false positives and performance degradation for such models [19], [25]. For instance, in Andersen et al. [5] the false positive rate increased from 1.7% to 4.5% when validating the model on an ambulatory database, which only contains ambulatory normal sinus rhythm (NSR) data from healthy individuals. Moreover, AF episodes occurrences will be rare, especially in the low AF burden population, and the noisy ambulatory recordings can often mimic such events. In our previous study [26], we showed that an AF detection DL model [5] trained on the AFDB dataset achieved excellent performance of around 98% accuracy. However, it resulted in a larger number of non-trivial false positive cases when applied on patient-operated ambulatory single channel ECG from free-living condition. In the same study, we found that nearly 62% of all the false positive cases were correlated with users ambulatory-context under free-living conditions. These false positive cases, which were mostly segments of length smaller than 50 seconds, were associated with three user contexts: (1) change in activity, (2) change in body position, and (3) sudden movement acceleration. It has been shown that such incorrect detection of AF in the longitudinal screening period could lead to over-diagnosis and patient anxiety [27].

Secondly, to reduce the complexity and achieve real-time detection, most of the AF classification models are primarily trained on the RRI based features without atrial activity analysis. Such models result in higher FPR in the presence of non-AF arrhythmias such as premature ventricular contractions (PVCs) and confounding noise in the ambulatory, which also exhibit irregular RRI characteristics similar to AF [5], [28], [29]. In a recent study, Tuboly et al. [29] also highlighted this problem. They showed that in the presence of non-AF arrhythmias, the numbers of false-positive AF detections are significantly higher if relied only on RRI features. Furthermore, Oster et al. [22] too pointed out that AF detection models trained only on RRI (heart's ventricular response) features are bound to result in high FPR on ECG from free-living conditions [22]. Jalali et al. [30] have tried to address the problem of AF misclassification due to the presence of PACs by using sensitivity and orthogonality constraints on a Residual Network (ResNet)'s cost function. They focused on detecting the irregularities before the AF onset that can indicate the onset of AF. Although this approach showed superior performance in the presence of PACs, the generality of such a model under-free living conditions with confounding noise and other artifacts remains unexplored.

In this paper, to reduce the FPR in longitudinal AF detection under free-living conditions and in the presence of confounding non-AF arrhythmias, we propose *DeepAware* – a hybrid multi-fusion based end-to-end AF detection model. The model is trained using both atrial activity and ventricular activity (i.e., RRI) based features. The proposed *DeepAware* model combines two of our previous algorithms as sub-models, which are introduced in [5] and [31]. Additionally, a context-aware heuristics (CAH) model is also developed in *DeepAware* model to analyzes patients' ambulatory contextual data. The context-aware heuristics module specifically enhances the AF detection results under free-living ambulatory conditions.

DeepAware model is validated successfully on several datasets, which shows its high generalizability. Moreover, the main differences between *DeepAware* and other state-of-the-art algorithms are the capability of the proposed model to reduce the number of false positive cases:

1) In the presence of many confounding non-AF arrhythmias.
2) Under free-living ambulatory conditions where confounding artifacts mimic AF.

It should be mentioned that the *DeepAware* model shows promising results in longitudinal AF screening under free-living conditions, which helps reduce the need to examine false-positive cases by physicians while preserving high sensitivity.

The remainder of this paper consists of 5 sections. Section II provides the methodology of the proposed algorithm. In Section III, the proposed *DeepAware* model is described in details. The results are presented and discussed in Section IV. Section V presents the limitations and future work, followed by the conclusion in Section VI.

II. Materials and Methods

A. Databases

In this study, six databases were used to train and validate the performance of the proposed model, which includes four PhysioNet databases such as MIT-BIH AF Database (AFDB) [32], QT database (QTDB) [33], MIT-BIH Arrhythmia Database (MITDB) [34], and two *in house* databases named CACHET Contextualised Arrhythmia Database (CACHET-CADB) and CACHET NSR Database (CACHET-NSRDB). Table I presents the technical specifications of these six databases.

The QTDB contains 105 recordings of 15 minutes each with a sampling frequency of 250 Hz, and the annotations include onset, peak, and offset labels of P, QRS, T, and U waves [33]. Likewise, AFDB includes 25 long-term ECG recordings of patients with paroxysmal AF. Each recording is nearly 10

TABLE I: Technical specifications of databases: Ch: No. of ECG channels, Freq: Sampling frequency, NS: Number of unique subjects in the recording, TR: Total number of records.

Database	Ch	Freq (Hz)	TR	Single Record Length	Total Duration	AF Duration h(%)	Unique Rhythms	NS	Contextual Data
AFDB	2	250	23	10h	234.3h	93.40 (39.87%)	4	25	✗
MITDB	2	360	48	0.5h	24.07h	2.16 (8.97%)	15	47	✗
NSRDB	2	128	18	24h	437.5h	0 (0%)	1	18	✗
CACHET-CADB	1	1024	1602	10sec	4.45h	2.07 (46.6%)	4	24	✓
CACHET-NSRDB	1	1024	10	24h	240 h	0 (0%)	1	10	✓
QTDB	2	250	105	15min	26.25h	n/a	n/a	105	✗

hours long and includes two channels of ECG collected at a sampling frequency of 250 Hz and 12-bit resolution [32], [35]. MITDB is sampled at 360 Hz and comprises 48 ECG records of 30 minutes long from 47 patients. Its annotations files include 14 different rhythms classes [34], [35]. On the other hand, MIT-BIH NSR Database (NSRDB) [35] contains 18 long-term two channels ECG from healthy subjects, which are mostly in NSR without any significant arrhythmias. The NSRDB is digitized at a sampling rate of 128 Hz.

The CACHET-CADB and CACHET-NSRDB are both *in house* databases, which contain patient operated single-channel contextualised ECG from free-living conditions. They are sampled at 1024 Hz and 12-bit resolution. Beside the ECG recordings, the patient's ambulatory contextual information such as activities, body positions, and movement accelerations are also provided with these two databases. These contextual data are obtained by processing the raw accelerometer, gyroscope, and pressure sensor of the chest-mounted ECG Holter for every 10 seconds interval. There are 1602 ECG records of 10 seconds length from 24 subjects in the CACHET-CADB. Each record belongs to one of the four classes, namely AF, NSR, noise, and 'others' class. On the other hand, CACHET-NSRDB contains 10 long-term normal sinus rhythm ECG records. Almost all the recordings in CACHET-NSRDB are 24 hours long. Both CACHET-CADB and CACHET-NSRDB were mainly used for evaluating the impact of context-aware heuristics (section III-E) on AF detection under free-living conditions.

B. Deep learning theory

Deep learning enables computational models to learn useful features directly from input data without any needed prior knowledge to engineer the features [36]. It has enhanced the state-of-the-art in domains such as image and speech recognition, natural language processing, drug discovery, and genomics [6], [36]. In recent years, deep learning has been successfully applied for the detection of AF and other type of arrhythmias [19], [37], [38].

The proposed *DeepAware* algorithm in this study combines two deep learning models, which were introduced in our previous studies [5], [31]. The first model is a combination of CNN and long-short term memory (LSTM) layers, which takes the heart rate variability (ventricular response) i.e., RR-intervals as inputs [5], [39]. It is denoted as RR-Net in the Fig. 2. The seconds model is likewise a combination of CNN and LSTM layers, which is used for p-wave detection from heart beats. The second model is called DENS-ECG [31]. As shown in Fig. 2, there are three layers of CNN followed by one LSTM layer. DENS-ECG takes raw ECG signals and outputs the number of detected p-waves. The output of these two models is combined with the context-aware heuristic model to detect the AF rhythms.

1) CNN layer: The CNN [40], [41] have been proven very effective in pattern recognition tasks. CNNs are capable of exploiting both spatial and temporal patterns in the data [36]. To achieve this, CNNs follow four key ideas, which are: 1) local connections; 2) shared weights for convolution process; 3) create large number of filters; and 4) reduce the network complexity as much as possible. Besides input and output layers a typical CNN structures consist of one or more connected convolutional layers, pooling layers, ReLU, and normalization layers. Fig 1 depicts the CNN structure with input, convolution, and pooling layers. In 1D-CNNs for analyzing ECG signals, various filters are generated by sliding a fixed window over the ECG record. This process is called convolution and the size of window is known as kernel size (k_{size}). The weights of these kernels and the overall bias is to be learned during the training process. It should be noted that the weights of the kernel are fixed for each filter map [42].

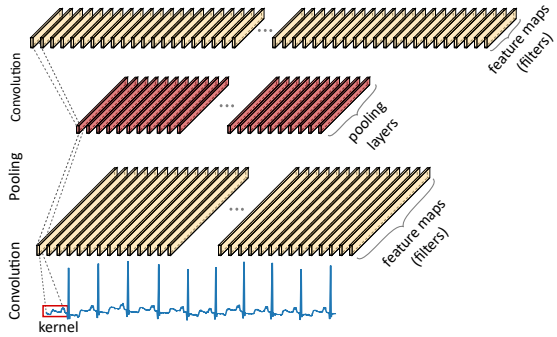


Fig. 1: CNN structure with input, convolution, and pooling layers

2) LSTM layer: Recurrent Neural Networks (RNNs) are specially designed to handle sequential and time-series data. RNNs can capture dependencies in sequential information very efficiently. However, it has been shown in the literature that learning long-term dependencies are very challenging [43]. On the other hand, the problem of unstable gradient can be solved by LSTM networks, which are a special type of RNNs. LSTM networks can handle long-term dependencies [44]. As shown in Figure 3, a LSTM block has three main parts: 1) input

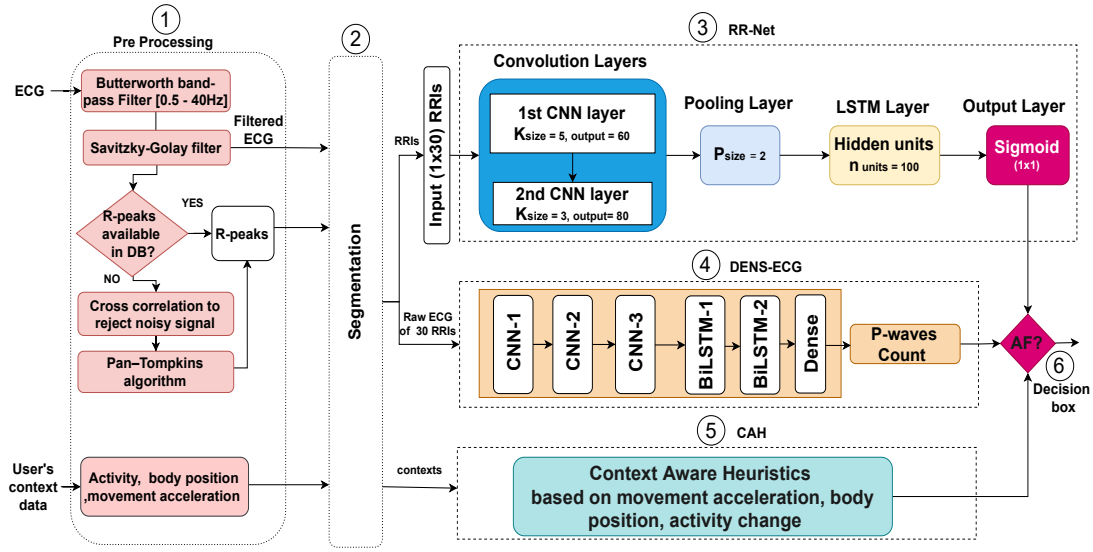


Fig. 2: The architecture of the proposed *DeepAware* model consists of six sub-components: (1) ECG data preprocessing, (2) segmentation, (3) the RR-Net, which take inputs of the RR interval series, (4) the *DENS-ECG*, which takes the raw ECG inputs and gives P-wave count, (5) and a CAH model, which takes user context in a case of ambulatory ECG for dynamically fine-tuning the final output, and (6) AF decision box for final binary output.

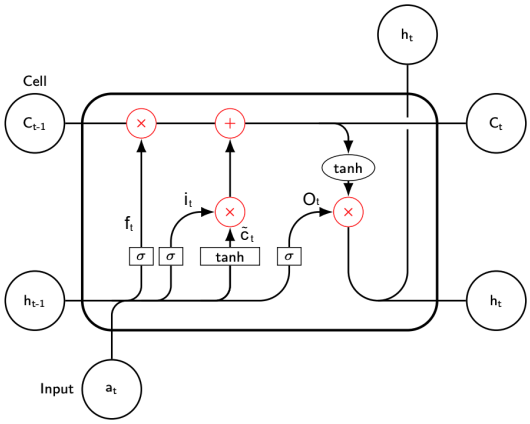


Fig. 3: LSTM memory block

gate (i_t), 2) forget gate (f_t), and 3) output gate (o_t). Forget and input gates control the flow of information removal and addition to the memory block as follows:

$$f_{(t)} = \sigma(\mathbf{u}_f^T \mathbf{a}_t + \mathbf{w}_f^T \mathbf{h}_{t-1} + b_t), \quad (1)$$

$$i_{(t)} = \sigma(\mathbf{u}_i^T \mathbf{a}_t + \mathbf{w}_i^T \mathbf{h}_{t-1} + b_t), \quad (2)$$

where $\mathbf{a}_t$ is the output from the previous layer and is the input sequence to the LSTM block at time step t, and $\mathbf{h}_{t-1}$ is the output sequence of the LSTM block at time $t-1$. The trainable parameters of the LSTM block are $\mathbf{w}_f$, $\mathbf{u}_f$, $\mathbf{w}_i$, $\mathbf{u}_i$, b_f, and b_i, which are weight vectors and bias terms. The memory of a LSTM block, c_t, is updated as follows:

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t, \quad (3)$$

where

$$\tilde{c}_t = \tanh(b_c + \mathbf{u}_c^T \mathbf{a}_t + \mathbf{w}_c^T \mathbf{h}_{t-1}). \quad (4)$$

Consequently, the output of the LSTM block is generated by:

$$h_t = o_t \tanh(c_t), \quad (5)$$

where

$$o_t = \sigma(\mathbf{w}_o^T \mathbf{a}_t + \mathbf{u}_o^T \mathbf{h}_{t-1} + b_o). \quad (6)$$

In (6), $\mathbf{u}_o$ and $\mathbf{w}_o$ are the weight vectors and b_o is the bias of the output gate. It can be seen from (5) and (6) that LSTM is capable of keeping or forgetting the existing memory efficiently [45].

Bidirectional LSTM is a variant of LSTM, which unlike LSTM can process the sequential time-series in both forward and backward directions with two separate hidden layers. Bidirectional LSTM have been found very useful in several ECG classification algorithms [5], [46]. In our proposed *DeepAware* algorithm, bidirectional LSTM has been used in both DENS-ECG and RR-Net models (Figure 2).

III. DeepAware Architecture

Fig 2 illustrates the flowchart of the proposed *DeepAware* algorithm. It comprises of six main components: (1) ECG data preprocessing, (2) segmentation, (3) RR-Net, (4) DENS-ECG model for the p-wave count, (5) the context-aware heuristic model, and (6) the AF decision box. In this section, all these six components will be described in detail.

A. Data preparation and pre-processing

As shown in Figure 2, data preparation and pre-processing is the first step. The baseline wanders and high-frequency noises were removed using a band-pass (0.5–40 Hz) filter. Then, the ECG signals were smoothed by applying a Savitsky-Golay filter [47]. It should be noted that, in some databases like AFDB, MITDB, and NSRDB, the R-peak locations are already available within the database, whereas, for CACHET-CADB and CACHET-NSRDB, the Pan–Tompkins algorithm [48] was used for finding R-peaks locations.

B. Segmentation

The RRIs and the filtered ECG signal obtained in the previous step are segmented into a window length of 30 RRIs. The sliding window has an overlap of 10 RRIs. The segmented windows (30 RRIs) are provided as input to the RR-Net model. The corresponding ECG segments, which has the same size as 30 RRIs, was fed simultaneously as input to the DENS-ECG model. Similarly, the CACHET-CADB and CACHET-NSRDB databases, which include the information about the user's context such as activity, body position, and movement acceleration data, were also segmented into time-duration equal to that of 30 RRIs.

C. RR-Net

Irregular RRIs is considered as one of the strong indications of AF in ECGs. The RR-Net model takes the windows of RRIs of size 1×30 as inputs. The model is a combination of two convolutional layer followed by a LSTM layer [5]. The convolutional layers extract the features from the RRIs, which are used by the LSTM layer afterwards. The first convolutional layer uses a kernel of size 5 ($K_{size} = 5$) and outputs 60 features. The input sequences are zero-padded to preserve the temporal dimensionality. The outputs of the first convolutional layer are considered as inputs to the second convolution layer, which has a kernel size of 3 ($K_{size} = 3$). The second convolutional layer generates more abstract features. Similar to the first layer, zero-padding is also applied to preserve the temporal dimensionality. As depicted in Figure 2, a max pooling layer is applied after the two convolutional layers, which has a kernel size of 2 ($P_{size} = 2$) with strides of two. The pooling layer results in reducing the temporal dimension of the inputs by half, which is an essential step for bringing down the complexity before the LSTM layer. The output of the pooling layer is fed into the bidirectional LSTM layer consists of $n_{units} = 100$ hidden units. The output of LSTM is fed into the classification layer with a sigmoid neuron. The sigmoid neuron's output can be considered a posterior probability of the degree of irregularity for the i'th RRIs (input sequence). These probabilities is finally converted to a binary output of RR-Net model as follows:

$$\text{RR-Net}(i) = \begin{cases} 1, & \text{if } p(y_i = irregular | \, \mathbf{x}_i, \text{RR-Net}) \geq 0.5, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where RR-Net(i) is the predicted irregularity for the i'th RRIs segment and the probability threshold is set to 0.5. As shown in Figure 2, the output of RR-Net is further used by the decision box.

D. DENS-ECG model

The DENS-ECG model is a combination of three convolutional layers and a dropout layer followed by two BiLSTM layers [31]. The 1D convolutional layers extract high abstract features from ECG segments. The two deep LSTM layers is used to process the extracted features by the previous 1D convolutional layers. The three convolutional layers use a kernel size of 3 ($K_{size} = 3$) and the number of filters (feature maps) for the three successive layers are 32, 64, and 128, respectively. In addition, zero padding is applied to maintain the same dimension in the input and convolutional layers. For example, the output of the third convolutional layer is 128 feature maps, which are then used as inputs for the first LSTM layer. The corresponding number of hidden units (n_{units}) are 250 and 125 for the two LSTM layers, respectively. Finally, The output of the second LSTM layer is fed into a dense layer, which generate posterior probabilities for the P-, QRS, T-, and No-wave segments of the ECG signals. As presented in Figure 2, the number of the P-waves detected by DENS-ECG model is provided to the decision box. It is worth noting that the dropout layer after the third convolutional layer helps preventing the over-fitting problem during the training phase of the model. The dropout probability is set to 0.2, which means 20% of the units is set to zero at each training step.

The absence of P-waves for the i'th ECG segments is computed as follows:

$$\text{DENS-ECG}(i) = \begin{cases} 1, & \text{If } P_{counts} \leq 15 \text{ for 31 heartbeats (30 RRIs)}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where DENS-ECG(i) is the predicted P-wave for the i'th RRIs segment and P_{counts} is the number of P-waves detected by the DENS-ECG model for the i'th ECG segment as the threshold is set to 15. As shown in Figure 2, the output of DENS-ECG is further used by the decision box.

E. Context-Aware Heuristics

The context-aware heuristics (CAH) model is based on our previous work [26] in which we analyzed the relationship between the false positive rate and user's context. Further analysis of false-positive cases using contextual data concluded that the vast majority ($\sim$99%) of false-positive cases, which have episodes shorter than 50 seconds, were associated with three main contexts: 1) change of activity; 2) change in body position (especially during laying/sleep); and 3) sudden movement acceleration.

The CAH model evaluates if there is a change in user's context during a specific 30 RRIs segment or its preceding segment. As shown in Figure 2, the CAH model assigns a binary output to detect whether a context change is detected with the current or previous RRI input window. Any identified

context changes result in a non-AF episode detection for the corresponding RRIs segment, which is done as follows:

$$\text{CAH}(i) = \begin{cases} 0, & \text{if context change detected,} \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

where $\text{CAH}(i)$ is the prediction for the i'th RRIs segment.

F. The decision box

As depicted Figure 2, the outputs of the three models (RR-Net, DENS-ECG, and Context-Aware Heuristics) are combined using the decision box. This is performed as follows:

$$\widehat{D}(i) = \begin{cases} \text{RR-Net}(i) \wedge \text{DENS-ECG}(i), \text{if context not available,} \\ \text{RR-Net}(i) \wedge \text{DENS-ECG}(i) \wedge \text{CAH}(i), \text{otherwise,} \end{cases} \tag{10}$$

where $\widehat{D}(i)$ is the final binary classification for the ith input sequence.

IV. Results and Discussion

Two publicly available datasets, the AFDB and QTDB were used for training the RR-Net and DENS-ECG sub-models, respectively. First, a 10-fold cross validation technique [53] is applied to train the model. Then, it is evaluated on two public (MITDB and NSRDB) and two *in house* datasets (CACHET-CADB and CACHET-NSRDB). The details of these datasets are available in Table I. Specifically, the MITDB dataset consists of 13 other types of arrhythmias. The performance of the model on MITDB indicate its generalizability in the presence of PVCs and other non-AF arrhythmias. On the other hand, both NSRDB and CACHET-NSRDB datasets only contain normal rhythms, which are used to evaluate model's performance and expected FPR on healthy subjects. In addition, the CACHET-NSRDB and CACHET-CADB datasets contain user's contextual information during the ambulatory ECG under free-living conditions. These two datasets are specifically used to examine the effectiveness of the context-aware heuristics model in reducing the FPs induced by the change in user ambulatory contexts.

All the individual sub-modules of *DeepAware* (Figure 2) were built in python 3.7 using Tensorflow 2.4.1 framework. The training was done on macOS 10.15.7 with 16 GB RAM, Dual-Core Intel Core i7 processed and an Intel Iris Plus Graphics 650 1536 MB graphics card. The metrics used for evaluating the performance of *DeepAware* and the obtained results on each dataset are described in the following sections.

A. Model Evaluation Metrics

Classification systems are usually evaluated by their capability of correctly classifying new samples. The performance of the binary classification problems are evaluated using a confusion matrix as given in Table III.

Average accuracy is one of the most commonly used metrics to evaluate the performance of classification models, which can be calculated as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{11}$$

In addition, other well-known metrics can be derived from Table III to report the performance of classifiers, which are given as follows:

$$Se = \frac{TP}{TP + FN}. \tag{12}$$

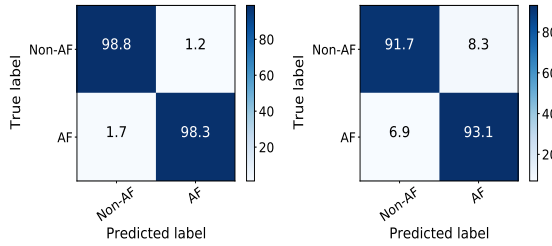
$$Sp = \frac{TN}{TN + FP}. \tag{13}$$

$$FPR = \frac{FP}{FP + TN}. \tag{14}$$

where Se, Sp, and FPR are sensitivity, specificity, and false positive rate, respectively.

B. DeepAware performance on public datasets and Comparison with the state-of-the-art

Table II compares the performance of the proposed *DeepAware* model with other state-of-the-art models on the AFDB, MITDB, and NSRDB datasets. Figure 4a and 4b also show the confusion matrices of the proposed *DeepAware* model on AFDB and MITDB datasets, respectively. The *DeepAware* model clearly outperforms the state-of-the-art algorithms on AFDB and it is generalized enough to perform well on unseen datasets such as MITDB and NSRDB. It archived a sensitivity, specificity and accuracy of 98.27%, 98.84%, and 98.62%, respectively, on AFDB dataset using a 10-fold cross-validation.



(a) (b)

Fig. 4: Confusion matrix of: (a) AFDB, (b) MITDB. The numbers are in percentage

The proposed model also achieved a sensitivity, specificity and accuracy of 93.05%, 91.67%, and 91.82% on MITDB dataset, respectively. The proposed *DeepAware* model improved the reported specificity and accuracy in [5] on MITDB dataset by 5.63% and 4.42%, respectively, at the cost of 5.91% reduction in sensitivity. It should be mentioned that the performance of the *DeepAware* model on MITDB dataset indicates its robustness in the presence of PVC/VPC beats and non-AF arrhythmias. The ectopic beats and non-AF arrhythmia resemble the AF in terms of irregularity in the RR intervals, thereby causing more FPs in AF detection models [29], [54], [55]. For example, in [5] (see Table II), despite high specificity on AFDB dataset, model's specificity on the MITDB, which has 14 other types of non-AF arrhythmias, has reduced drastically to 86.04%.

TABLE II: Comparison of *Deep Aware* algorithm with other state-of-the-art models on AFDB, MITDB, and NSRDB datasets. All the results are in percentage.

Algorithm	Methods	Features	Ch	AFDB				MITDB				NSRDB	
				Se	Sp	Acc	FPR	Se	Sp	Acc	FPR	Sp	FPR
[49]	CNN, BLSTM	RRI, Heartbeat Sequences	1	99.93	97.03	96.59	-	-	-	-	-	-	-
[50]	CNN	MFSWT	1	74.96	86.41	81.07	-	-	-	-	-	-	-
[25]	CNN	SWT, STFT	1	98.79	97.87	98.63	-	-	-	-	-	-	-
[51]	CNN	RRI, F-wave frequency spectrum	1	97.4	96.2	97.3	-	-	-	-	-	-	-
[52]	CNN, RCN	Raw ECG	2	94.28	94.91	94.59	-	-	-	-	-	-	-
[18]	MCNN	Instant Heart Rate Sequence	2	98.22	98.11	98.18	-	-	-	-	-	-	-
[5]	CNN, BLSTM	RRI	2	98.17	96.29	97.1	3.71	98.96	86.04	87.4	13.96	95.01	4.99
DeepAware	CNN, BLSTM	RRI, Raw EEG, Context	1	98.27	98.84	98.62	1.16	93.05	91.67	91.82	8.33	98.47	1.53

Ch: Number of ECG channels; MFSWT: Modified Frequency Slice Wavelet Transform; MCNN: Multi-Scale CNN; IHR: Instant Heart Rate Sequence; RCN: Recurrence Complex Network

TABLE III: Confusion matrix

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

Furthermore, the *DeepAware* model has been generalized well to perform better on NSRDB dataset compared to other state-of-the-art models. As it can be seen in Table II, the proposed *DeepAware* model improved the specificity reported in [5] by 3.57%.

C. DeepAware performance on contextualised ECG datasets

The context-aware heuristics model was evaluated on two *in house* CACHET-CADB and CACHET-NASRDB datasets, which are patient-operated contextualized ECG datasets under free-living conditions. As reported in Table IV, the proposed *DeepAware* model has achieved a sensitivity, specificity, and accuracy of 97.94%, 98.39%, 98.06% on CACHET-CADB dataset, respectively. Figure 5 shows the confusion matrix of *DeepAware* model on CACHET-CADB dataset. Similarly, Table V reports the performance of the *DeepAware* model on CACHET-NSRDB dataset. The average FPR for all the records in Table V is 1.76%. As given in Table V, in general, the proposed *DeepAware* model outperforms the RR-Net model on CACHET-NSRDB dataset.

TABLE IV: Performance on CACHET-CADB.

Measure	CACHET-CADB	
	RR-Net	DeepAware
Se (%)	99.63	97.94
Sp (%)	90.32	98.39
Acc (%)	97.22	98.06
FPR (%)	9.68	1.61

Additionally, a comparison between the performance of RR-Net and *DeepAware* models in Table IV and Table V on

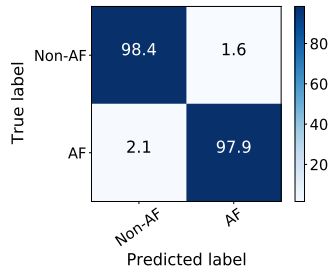


Fig. 5: Confusion matrix of CACHET-CADB. All number are in percentage.

CACHET-CADB and CACHET-NSRDB confirms the positive effect of the context-heuristics approach to lower down the FPR on patient-operated ECG under free-living conditions.

TABLE V: Performance on CACHET-NSRDB. Each record consists of 24 hours long contextualised ECG under free living conditions from healthy individuals.

Record	Inputs No.	R-peaks	RR-Net		DeepAware	
			Sp	FPR	Se	FPR
1	5714	114319	89.1	10.9	98.41	1.59
2	5906	118156	88.52	11.47	95.54	4.45
3	3998	80037	89.37	10.63	99.39	0.61
4	3535	70733	91.85	8.15	99.8	0.2
5	1429	28634	97.06	2.93	99.02	0.8
6	4123	82565	82.77	17.23	98.16	1.84
7	5388	108046	95.36	4.64	96.82	3.18
8	5959	119276	80.86	19.13	96.29	3.71
9	4600	92173	94.043	5.95	99.54	0.46
10	5017	100396	95.25	4.7	99.36	0.76

Inputs No.: Number of (30x1) input windows

On CACHET-CADB, as compared to RR-Net, *DeepAware* has improved the specificity and reduced the FPR by around 8% at the cost of a 1.69% reduction in the sensitivity. Similarly, on CACHET-NSRDB, the average FPR has been reduced

by 7.81%. Furthermore, as CACHET-NSRDB only contains subjects having normal sinus rhythms, the performance of *DeepAware* on CACHET-NSRDB is a good indication of expected FPR in healthy and low AF prevalence subjects under free-living conditions.

D. AF detection with and without atrial activity analysis on various datasets

Table VI compares the performance of RR-Net and *DeepAware* model on MITDB and NSRDB datasets, which signifies the impact of including atrial activity analysis (i.e P-wave detection using DENS-ECG model) in AF detection algorithms. These results are in consistent with the findings in [29], which also showed that taking atrial activity analysis features into account can reduce the false positive rate of RRI based models significantly.

As given in Table VI, including atrial activity feature in the proposed *DeepAware* model (using DENS-ECG model), reduced the FPR by 4.57%. It should be noted that respiratory sinus arrhythmia, which is a natural response of the healthy heart, can be misclassified as AF without the analysis of atrial activity. This false diagnosis usually occurs in the young population with a low prevalence of AF [29]. It is also important to highlight that existing literature on deep-learning-based AF detection has limited coverage of examining the impact of non-AF arrhythmias (i.g, sinus arrhythmia) on FPR and specificity of AF detection algorithms.

TABLE VI: Comparison of the classification performance between RR-Net and *DeepAware* model on MITDB and NSRDB datasets.

Measure	MITDB		NSRDB	
	RR-Net	DeepAware	RR-Net	DeepAware
Se [%]	97.74	93.06	-	-
Sp [%]	87.10	91.67	95.53	98.47
Acc [%]	88.22	91.82	-	-
FPR [%]	12.90	8.33	4.47	1.53

E. Performance improvement by context-aware heuristics model under free-living ambulatory conditions

Figure 6 shows a typical scenario of an ECG signal captured under free-living ambulatory conditions. The irregularity in RRI, which is induced by a change in the user's ambulatory contexts, may lead to an AF diagnosis. Such RRI irregularities on context change are either heart' natural response to change or could be due to motion artifacts. The irregularity induced by a change in the context is usually short (30-60 seconds) [26]. The context-aware heuristics part in the proposed *DeepAware* model helps identifying whether the RRI irregularity detected by RR-Net is in fact due to the heart diseases or it is just a change in the user's ambulatory context. The impact of the context-aware heuristics (CAH) model in reducing the FPR under free-living ambulatory conditions can be observed in both Table IV and Table V.

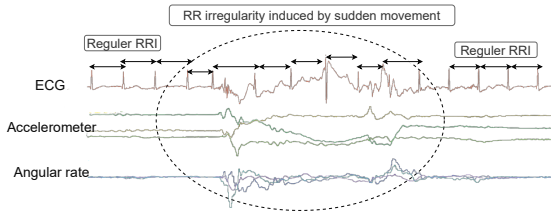


Fig. 6: An example of irregular RRI caused by changes in user's ambulatory context, which resembles an AF episode. The single lead ECG signal along with accelerometer and angular rate are shown.

V. Limitations and Future Work

The presented *DeepAware* model has two main limitations that require further improvements. First, compared to RRI based model such as RR-Net model, the proposed *DeepAware* model is computationally expensive. The RR-Net can classify 24 hours of ECG in $\leq$ 1 minute, whereas it takes more than 30 minutes for *DeepAware* to analyze the same amount of the data in a non-GPU computing environment. Therefore, it may not be straightforward to deploy this model in resource-constrained wearable devices. However, the *DeepAware* model will be more suitable to be used in a cloud computing environment.

Secondly, following the limitation of DENS-ECG model in detecting inverted P-waves, the *DeepAware* model may encounter more FPs in the presence of such ECG morphology with inverted P-waves. To overcome this, the DENS-ECG model can be trained on a dataset with higher number of inverted P-waves morphology, which is currently missing in QTDB [33].

VI. Conclusion

This article presented a hybrid end-to-end AF detection algorithm, named as, *DeepAware*, by combining DL and context-aware heuristics. The model receives three different inputs: (1) RRIs, (2) raw ECG signals, and (3) patient's ambulatory context, and it outputs the binary classification of AF and non-AF rhythms. Unlike most state-of-the-art, the *DeepAware* model has been evaluated on five different datasets, four of which are unseen to the model during the training phase. The proposed *DeepAware* model achieved better AF detection performance on public datasets than the state-of-the-art and, most importantly, on the patient-operated ambulatory mobile ECG from free-living conditions. The performance improvement by the context-aware heuristics model highlights the capability of context-awareness to enhance AF detection. We also demonstrated that only relying on RRI features for AF detection is problematic and leads to high false positive rate, especially in the presence of confounding arrhythmias (i.e., atrial flutter, PVCs, atrial sinus arrhythmias), and ambulatory motion artifacts (from context change). The *DeepAware* model can significantly reduce the workload requiring the manual verification of the false positives in a clinical setting for longitudinal screening.

REFERENCES

[1] A. S. Go, E. M. Hylek, K. A. Phillips, Y. Chang, L. E. Henault, J. V. Selby, and D. E. Singer, "Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the anticoagulation and risk factors in atrial fibrillation (atria) study," *Jama*, vol. 285, no. 18, pp. 2370–2375, 2001.

[2] S. Ramkumar, N. Nerlekar, D. D'Souza, D. J. Pol, J. M. Kalman, and T. H. Marwick, "Atrial fibrillation detection using single lead portable electrocardiographic monitoring: a systematic review and meta-analysis," *BMJ open*, vol. 8, no. 9, p. e024178, 2018.

[3] S. M. P. Dinakarrao, A. Jantsch, and M. Shafique, "Computer-aided arrhythmia diagnosis with bio-signal processing: A survey of trends and techniques," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–37, 2019.

[4] Z. Wu, X. Ding, G. Zhang, X. Xu, X. Wang, Y. Tao, and C. Ju, "A novel features learning method for ecg arrhythmias using deep belief networks," in *2016 6th International Conference on Digital Home (ICDH)*. IEEE, 2016, pp. 192–196.

[5] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Systems with Applications*, vol. 115, pp. 465–473, 2019.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten chinese character recognition using googlenet and directional feature maps," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 846–850.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[9] A. Akay and H. Hess, "Deep learning: Current and emerging applications in medicine and technology," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 906–920, 2019.

[10] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[11] M.-Z. Poh, Y. C. Poh, P.-H. Chan, C.-K. Wong, L. Pun, W. W.-C. Leung, Y.-F. Wong, M. M.-Y. Wong, D. W.-S. Chu, and C.-W. Siu, "Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms," *Heart*, vol. 104, no. 23, pp. 1921–1928, 2018.

[12] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2095–2104, 2018.

[13] O. Faust, A. Shenfield, M. Kareem, T. R. San, H. Fujita, and U. R. Acharya, "Automated detection of atrial fibrillation using long short-term memory network with rr interval signals," *Computers in biology and medicine*, vol. 102, pp. 327–335, 2018.

[14] J. Wang, "A deep learning approach for atrial fibrillation signals classification based on convolutional and modified elman neural network," *Future Generation Computer Systems*, vol. 102, pp. 670–679, 2020.

[15] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, p. 65, 2019.

[16] S. S. Xu, M.-W. Mak, and C.-C. Cheung, "Towards end-to-end ecg classification with raw signal extraction and deep neural networks," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1574–1584, 2018.

[17] M. Limam and F. Precioso, "Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[18] Z. Yao, Z. Zhu, and Y. Chen, "Atrial fibrillation detection by multi-scale convolutional neural networks," in *2017 20th International Conference on Information Fusion (Fusion)*. IEEE, 2017, pp. 1–6.

[19] 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," *IEEE journal of biomedical and health informatics*, vol. 22, no. 6, pp. 1744–1753, 2018.

[20] G. Petmezas, K. Haris, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J. A. Rogers, A. K. Katsaggelos, and N. Maglaveras, "Automated atrial fibrillation detection using a hybrid cnn-lstm network on imbalanced ecg datasets," *Biomedical Signal Processing and Control*, vol. 63, p. 102194, 2021.

[21] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ecg arrhythmia classification," *Expert Systems with Applications: X*, p. 100033, 2020.

[22] J. Oster and G. D. Clifford, "Impact of the presence of noise on rr interval-based atrial fibrillation detection," *Journal of electrocardiology*, vol. 48, no. 6, pp. 947–951, 2015.

[23] H. Gao, C. Liu, X. Wang, L. Zhao, Q. Shen, E. Ng, and J. Li, "An open-access ecg database for algorithm evaluation of qrs detection and heart rate estimation," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 9, pp. 1853–1858, 2019.

[24] M. F. Márquez, L. Colín, M. Guevara, P. Iturralde, and A. G. Hermosillo, "Common electrocardiographic artifacts mimicking arrhythmias in ambulatory monitoring," *American heart journal*, vol. 144, no. 2, pp. 187–197, 2002.

[25] Y. Xia, N. Wulan, K. Wang, and H. Zhang, "Detecting atrial fibrillation by deep convolutional neural networks," *Computers in biology and medicine*, vol. 93, pp. 84–92, 2018.

[26] D. Kumar, S. Puthusserypady, H. Dominguez, K. Sharma, and J. E. Bardram, "Contextual and temporal distribution of false positives in a deep learning-based atrial fibrillation detection algorithm: An investigation," *Expert Systems with Applications*, Forthcoming.

[27] C. C. Cheung, A. D. Krahn, and J. G. Andrade, "The emerging role of wearable technologies in detection of arrhythmia," *Canadian Journal of Cardiology*, vol. 34, no. 8, pp. 1083–1087, 2018.

[28] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark, "Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[29] G. Tuboly, G. Kozmann, O. Kiss, and B. Merkely, "Atrial fibrillation detection with and without atrial activity analysis using lead-i mobile ecg technology," *Biomedical Signal Processing and Control*, vol. 66, p. 102462, 2021.

[30] A. Jalali and M. Lee, "Atrial fibrillation prediction with residual network using sensitivity and orthogonality constraints," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 407–413, 2019.

[31] A. Peimankar and S. Puthusserypady, "Dens-ecg: A deep learning approach for ecg signal delineation," *arXiv preprint arXiv:2005.08689*, 2020.

[32] G. Moody, "A new method for detecting atrial fibrillation using rr intervals," *Computers in Cardiology*, pp. 227–230, 1983.

[33] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg," in *Computers in cardiology 1997*. IEEE, 1997, pp. 673–676.

[34] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[35] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] U. R. Acharya, H. Fujita, S. L. Oh, U. Raghavendra, J. H. Tan, M. Adam, A. Gertych, and Y. Hagiwara, "Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network," *Future Generation Computer Systems*, vol. 79, pp. 952–959, 2018.

[38] D. Lai, Y. Bu, Y. Su, X. Zhang, and C.-S. Ma, "Non-standardized patch-based ecg lead together with deep learning based algorithm for automatic screening of atrial fibrillation," *IEEE journal of biomedical and health informatics*, vol. 24, no. 6, pp. 1569–1578, 2020.

[39] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580–4584.

[40] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[41] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[42] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[43] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[45] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[46] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification," *Computers in biology and medicine*, vol. 96, pp. 189–202, 2018.

[47] W. H. Press and S. A. Teukolsky, "Savitzky-golay smoothing filters," *Computers in Physics*, vol. 4, no. 6, pp. 669–672, 1990.

[48] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.

[49] H. Dang, M. Sun, G. Zhang, X. Qi, X. Zhou, and Q. Chang, "A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals," *IEEE Access*, vol. 7, pp. 75 577–75 590, 2019.

[50] X. Xu, S. Wei, C. Ma, K. Luo, L. Zhang, and C. Liu, "Atrial fibrillation beat identification using the combination of modified frequency slice wavelet transform and convolutional neural networks," *Journal of healthcare engineering*, vol. 2018, 2018.

[51] D. Lai, X. Zhang, Y. Zhang, and M. B. B. Heyat, "Convolutional neural network based detection of atrial fibrillation combing rr intervals and f-wave frequency spectrum," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 4897–4900.

[52] X. Wei, J. Li, C. Zhang, M. Liu, P. Xiong, X. Yuan, Y. Li, F. Lin, and X. Liu, "Atrial fibrillation detection by the combination of recurrence complex network and convolution neural network," *Journal of Probability and Statistics*, vol. 2019, 2019.

[53] D. Berrar, "Cross-validation," *Encyclopedia of bioinformatics and computational biology*, vol. 1, pp. 542–545, 2019.

[54] J. Selder, L. Breukel, S. Blok, A. van Rossum, I. Tulevski, and C. Allaart, "A mobile one-lead ecg device incorporated in a symptom-driven remote arrhythmia monitoring program. the first 5,982 hartwacht ecgs," *Netherlands Heart Journal*, vol. 27, no. 1, pp. 38–45, 2019.

[55] P.-H. Chan, C.-K. Wong, L. Pun, Y.-F. Wong, M. M.-Y. Wong, D. W.-S. Chu, and C.-W. Siu, "Head-to-head comparison of the alivecor heart monitor and microlife watchbp office afib for atrial fibrillation screening in a primary care setting," *Circulation*, vol. 135, no. 1, pp. 110–112, 2017.

APPENDIX B

List of CUMACF Questionnaires in mCardia's usability and feasibility study

List of CUMACF (CACHET Unified Methodology for Assessment of Clinical Feasibility) Questions used	
	Questions
Q1	Overall, I would find the system useful in home based longitudinal ECG collection for arrhythmia screening
Q2	I would use mCardia daily basis as instructed
Q3	Using mCardia would increases the quality of communication b/w me and my doctor
Q4	Using mCardia would reduce recall bias in reporting my symptoms during screening period
Q5	mCardia would help me in keeping track of my daily activeness and unusual symptoms and help me understand my symptoms better
Effort Expectancy	
Q6	Overall, I would be satisfied with how easy it is to use mCardia App
Q7	My interaction with mCardia would be clear and understandable.
Q8	It would be easy for me to learn to use mCardia App
Q9	I would find mCardia easy to use
Q10	I would be skillful at using mCardia
Q11	The information (such as [error messages | help | messages | guidelines | tutorials | …]) provided with mCardia are clear and useful
Q12	The interface was effective in helping me complete the task [events entry]
Q13	mCardia was pleasant to use
Q14	mCardia has all the functionalities that I expect it to have
Social Influence	
Q15	My doctor thinks that I should use mCardia
Q16	My family [spouse | children | parents | ...] think that I should use mCardia
Facilitating Conditions	
Q17	I have the resources necessary to use mCarida app
Q18	I have the knowledge necessary to use mCardia app
Q19	A specific person should be available for assistance with mCardia if I face any difficulty with mCardia App

Bibliography

[1] About arrhythmia. *Available [Online]: https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia*, Sep 2016.

[2] Cardiovascular diseases (cvds). *Available [Online]: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*, May 2017.

[3] U Rajendra Acharya, Hamido Fujita, Oh Shu Lih, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Automated detection of arrhythmias using different intervals of tachycardia ecg segments with convolutional neural network. *Information sciences*, 405:81–90, 2017.

[4] U. Rajendra Acharya, Hamido Fujita, Shu Lih Oh, U. Raghavendra, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Yuki Hagiwara. Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. *Future Generation Computer Systems*, 79:952–959, 2018.

[5] E. R. Adams and A. Choi. Using neural networks to predict cardiac arrhythmias. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 402–407, 2012.

[6] A. S. R. M. Ahouandjinou, K. Assogba, and C. Motamed. Smart and pervasive icu based-iot for improving intensive health care. In *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*, pages 1–4, 2016.

[7] Khalid K. Al-jabery, Tayo Obafemi-Ajayi, Gayla R. Olbricht, and Donald C. Wunsch II. 4 - selected approaches to supervised learning. In Khalid K. Al-jabery, Tayo Obafemi-Ajayi, Gayla R. Olbricht, and Donald C. Wunsch II, editors, *Computational Learning Approaches to Data Analytics in Biomedical Applications*, pages 101–123. Academic Press, 2020.

[8] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, 2017.

[9] A. Aliamiri and Y. Shen. Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 442–445, 2018.

[10] Rasmus S Andersen, Abdolrahman Peimankar, and Sadasivan Puthusserypady. A deep learning approach for real-time detection of atrial fibrillation. *Expert Systems with Applications*, 115:465–473, 2019.

[11] Daejoon Anh, Subramaniam Krishnan, and Frank Bogun. Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis. *Journal of electrocardiology*, 39(3):343–345, 2006.

[12] J Augusto, Asier Aztiria, Dean Kramer, and Unai Alegre. A survey on the evolution of the notion of context-awareness. *Applied Artificial Intelligence*, 31(7-8):613–642, 2017.

[13] Eric Awtry, Cathy Jeon, and Molly G Ware. *Blueprints cardiology*. Lippincott Williams & Wilkins, 2006.

[14] Jakob E Bardram. The art of doing a phd. Homepage, June 2016. Retrieved July 01, 2019 from `http://www.bardram.net/wp-content/uploads/2017/12/Art.PhD_.DTU_.v.30.pdf`.

[15] Jakob E. Bardram. CACHET Unified Methodology for Assessment of Clinical Feasibility. Technical report, Copenhagen Center for Health Technology, Copenhagen , Denmark, 2018. Available from http://www.cachet.dk/research/cumacf.

[16] Jakob E. Bardram. Cachet unified method for assessment of clinical feasibility, April 2019.

[17] Jakob E. Bardram. The carp mobile sensing framework – a cross-platform, reactive, programming framework and runtime environment for digital phenotyping, 2020.

[18] Jakob E. Bardram. Carp mobile sensing framework in flutter, 2020.

[19] Jakob E Bardram. Evaluating personalized pervasive health technology—but how? *IEEE Annals of the History of Computing*, 19(03):37–44, 2020.

[20] Syed Khairul Bashar, Dong Han, Shirin Hajeb-Mohammadalipour, Eric Ding, Cody Whitcomb, David D McManus, and Ki H Chon. Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches. *Scientific reports*, 9(1):1–10, 2019.

[21] Maria A Baturova, Arne Lindgren, Jonas Carlson, Yuri V Shubik, S Bertil Olsson, and Pyotr G Platonov. Predictors of new onset atrial fibrillation during 10-year follow-up after first-ever ischemic stroke. *International journal of cardiology*, 199:248–252, 2015.

[22] Noureddine Belgacem and Selma Boumerdassi. Mobile personal electrocardiogram monitoring system with patient location. In *Proceedings of the 1st ACM international workshop on Medical-grade wireless networks*, pages 69–72. ACM, 2009.

[23] Frank Bogun, Daejoon Anh, Gautham Kalahasty, Erik Wissner, Chadi Bou Serhal, Rabih Bazzi, W Douglas Weaver, and Claudio Schuger. Misdiagnosis of atrial fibrillation and its clinical consequences. *The American journal of medicine*, 117(9):636–642, 2004.

[24] Robert O Bonow, Douglas L Mann, Douglas P Zipes, and Peter Libby. *Braunwald's heart disease e-book: A textbook of cardiovascular medicine*. Elsevier Health Sciences, 2011.

[25] Xin-Cheng Cao, Bin Yao, and Bin-Qiang Chen. Atrial fibrillation detection using an improved multi-scale decomposition enhanced residual convolutional neural network. *IEEE Access*, 7:89152–89161, 2019.

[26] Bhavin Chamadiya, Kunal Mankodiya, Manfred Wagner, and Ulrich G Hofmann. Textile-based, contactless ecg monitoring for non-icu clinical settings. *Journal of Ambient Intelligence and Humanized Computing*, 4(6):791–800, 2013.

[27] Pak-Hei Chan, Chun-Ka Wong, Yukkee C Poh, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Ming-Zher Poh, Daniel Wai-Sing Chu, and Chung-Wah Siu. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association*, 5(7):e003428, 2016.

[28] Christopher C Cheung, Andrew D Krahn, and Jason G Andrade. The emerging role of wearable technologies in detection of arrhythmia. *Canadian Journal of Cardiology*, 34(8):1083–1087, 2018.

[29] Armand Chocron, Julien Oster, Shany Biton, Franck Mendel, Meyer Elbaz, Yehoshua Zeevi, and Joachim Behar. Remote atrial fibrillation burden estimation using deep recurrent neural network. *IEEE Transactions on Biomedical Engineering*, 2020.

[30] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[31] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[32] Mayo Clinic. Atrial fibrillation - symptoms and causes, 2019.

[33] Douglas A Coast, Richard M Stern, Gerald G Cano, and Stanley A Briller. An approach to cardiac arrhythmia analysis using hidden markov models. *IEEE Transactions on biomedical Engineering*, 37(9):826–836, 1990.

[34] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[35] Hao Dang, Muyi Sun, Guanhong Zhang, Xingqun Qi, Xiaoguang Zhou, and Qing Chang. A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals. *IEEE Access*, 7:75577–75590, 2019.

[36] Aras R Dargazany, Paolo Stegagno, and Kunal Mankodiya. Wearabledl: Wearable internet-of-things and deep learning for big data analytics—concept, literature, and future. *Mobile Information Systems*, 2018, 2018.

[37] S Dash, KH Chon, S Lu, and EA Raeder. Automatic real time detection of atrial fibrillation. *Annals of biomedical engineering*, 37(9):1701–1709, 2009.

[38] Alan Davies and Alwyn Scott. *ECG Basics*, pages 19–33. Springer London, London, 2014.

[39] Philip De Chazal, Maria O'Dwyer, and Richard B Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE transactions on biomedical engineering*, 51(7):1196–1206, 2004.

[40] Sai Manoj Pudukotai Dinakarrao, Axel Jantsch, and Muhammad Shafique. Computer-aided arrhythmia diagnosis with bio-signal processing: A survey of trends and techniques. *ACM Computing Surveys (CSUR)*, 52(2):23, 2019.

[41] Marcus Dörr, Vivien Nohturfft, Noé Brasier, Emil Bosshard, Aleksandar Djurdjevic, Stefan Gross, Christina J Raichle, Mattias Rhinisperger, Raphael Stöckli, and Jens Eckstein. The watch af trial: Smartwatches for detection of atrial fibrillation. *JACC: Clinical Electrophysiology*, 5(2):199–208, 2019.

[42] Barbara J Drew, Robert M Califf, Marjorie Funk, Elizabeth S Kaufman, Mitchell W Krucoff, Michael M Laks, Peter W Macfarlane, Claire Sommargren, Steven Swiryn, and George F Van Hare. Practice standards for electrocardiographic monitoring in hospital settings: an american heart association scientific statement from the councils on cardiovascular nursing, clinical cardiology, and cardiovascular disease in the young: endorsed by the international society of computerized electrocardiology and the american association of critical-care nurses. *Circulation*, 110(17):2721–2746, 2004.

[43] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, page 100033, 2020.

[44] Linda M Eerikäinen, Alberto G Bonomi, Lukas RC Dekker, Rik Vullings, and Ronald M Aarts. Atrial fibrillation monitoring with wrist-worn photoplethysmography-based wearables: State-of-the-art review. *Cardiovascular Digital Health Journal*, 1(1):45–51, 2020.

[45] Estrella Everss-Villalba, Francisco Manuel Melgarejo-Meseguer, Manuel Blanco-Velasco, Francisco Javier Gimeno-Blanes, Salvador Sala-Pla, José Luis Rojo-Álvarez, and Arcadi García-Alberola. Noise maps for quantitative and clinical severity towards long-term ecg monitoring. *Sensors*, 17(11):2448, 2017.

[46] Xiaomao Fan, Qihang Yao, Yunpeng Cai, Fen Miao, Fangmin Sun, and Ye Li. Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ecg recordings. *IEEE journal of biomedical and health informatics*, 22(6):1744–1753, 2018.

[47] Oliver Faust, Edward J Ciaccio, and U Rajendra Acharya. A review of atrial fibrillation detection methods as a service. *International Journal of Environmental Research and Public Health*, 17(9):3093, 2020.

[48] Oliver Faust, Alex Shenfield, Murtadha Kareem, Tan Ru San, Hamido Fujita, and U Rajendra Acharya. Automated detection of atrial fibrillation using long short-term memory network with rr interval signals. *Computers in biology and medicine*, 102:327–335, 2018.

[49] Oliver Faust, Alex Shenfield, Murtadha Kareem, Tan Ru San, Hamido Fujita, and U. Rajendra Acharya. Automated detection of atrial fibrillation using long short-term memory network with rr interval signals. *Computers in Biology and Medicine*, 102:327–335, 2018.

[50] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[51] Josef Finsterer, Claudia Stöllberger, and Edmund Gatterer. Oral anticoagulation for ecg tremor artefact simulating atrial fibrillation. *Acta cardiologica*, 58(5):425–429, 2003.

[52] Abdur Forkan, Ibrahim Khalil, and Zahir Tari. Context-aware cardiac monitoring for early detection of heart diseases. In *Computing in Cardiology 2013*, pages 277–280. IEEE, 2013.

[53] Abdur Rahim Mohamad Forkan and Weichich Hu. A context-aware, predictive and protective approach for wellness monitoring of cardiac patients. In *2016 Computing in Cardiology Conference (CinC)*, pages 369–372. IEEE, 2016.

[54] Alessio Galli, Francesco Ambrosini, and Federico Lombardi. Holter monitoring and loop recorders: from research to clinical practice. *Arrhythmia & electrophysiology review*, 5(2):136, 2016.

[55] Hongxiang Gao, Chengyu Liu, Xingyao Wang, Lina Zhao, Qin Shen, EYK Ng, and Jianqing Li. An open-access ecg database for algorithm evaluation of qrs detection and heart rate estimation. *Journal of Medical Imaging and Health Informatics*, 9(9):1853–1858, 2019.

[56] David J. Gladstone, Melanie Spring, Paul Dorian, Val Panzov, Kevin E. Thorpe, Judith Hall, Haris Vaid, Martin O'Donnell, Andreas Laupacis, Robert Côté, Mukul Sharma, John A. Blakely, Ashfaq Shuaib, Vladimir Hachinski, Shelagh B. Coutts, Demetrios J. Sahlas, Phil Teal, Samuel Yip, J. David Spence, Brian Buck, Steve Verreault, Leanne K. Casaubon, Andrew Penn, Daniel Selchen, Albert Jin, David Howse, Manu Mehdiratta, Karl Boyle, Richard Aviv, Moira K. Kapral, and Muhammad Mamdani. Atrial fibrillation in patients with cryptogenic stroke. *New England Journal of Medicine*, 370(26):2467–2477, 2014. PMID: 24963566.

[57] David J. Gladstone, Rolf Wachter, Katharina Schmalstieg-Bahr, F. Russell Quinn, Eva Hummers, Noah Ivers, Tamara Marsden, Andrea Thornton, Angie Djuric, Johanna Suerbaum, Doris von Grünhagen, William F. McIntyre, Alexander P. Benz, Jorge A. Wong, Fatima Merali, Sam Henein, Chris Nichol, Stuart J. Connolly, Jeff S. Healey, SCREEN-AF Investigators, and Coordinators. Screening for Atrial Fibrillation in the Older Population: A Randomized Clinical Trial. *JAMA Cardiology*, 02 2021.

[58] Movisens GmbH. Dataanalyzer-sensor data analysis, December 2020.

[59] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[60] Alex Graves. Long short-term memory. In *Supervised sequence labelling with recurrent neural networks*, pages 37–45. Springer, 2012.

[61] Scott David Greenwald. *The development and analysis of a ventricular fibrillation detector*. PhD thesis, Massachusetts Institute of Technology, 1986.

[62] Christopher A Groh, Madelaine Faulkner, Shiffen Getabecha, Victoria Taffe, Gregory Nah, Kathi Sigona, Debbe McCall, Mellanie True Hills, Kathleen Sciarappa, Mark J Pletcher, et al. Patient-reported triggers of paroxysmal atrial fibrillation. *Heart rhythm*, 16(7):996–1002, 2019.

[63] Jan Gulliksen, Bengt Göransson, Inger Boivie, Stefan Blomkvist, Jenny Persson, and Åsa Cajander. Key principles for user-centred systems design. *Behaviour and Information Technology*, 22(6):397–409, 2003.

[64] Hesam Halvaei, Emma Svennberg, Leif Sörnmo, and Martin Stridh. False alarm reduction in atrial fibrillation screening. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

[65] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.

[66] Anders Hansson, Bjarne Madsen-Härdig, and S Bertil Olsson. Arrhythmia-provoking factors and symptoms at the onset of paroxysmal atrial fibrillation: a study based on interviews with 100 patients seeking hospital assistance. *BMC cardiovascular disorders*, 4(1):13, 2004.

[67] Robert G Hart and Jonathan L Halperin. Atrial fibrillation and thromboembolism: a decade of progress in stroke prevention. *Annals of Internal Medicine*, 131(9):688–695, 1999.

[68] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[70] Runnan He, Kuanquan Wang, Na Zhao, Yang Liu, Yongfeng Yuan, Qince Li, and Henggui Zhang. Automatic detection of atrial fibrillation based on continuous wavelet transform and 2d convolutional neural networks. *Frontiers in physiology*, 9:1206, 2018.

[71] Jeffrey S. Healey and Roopinder K. Sandhu. Are we ready for mass screening to detect atrial fibrillation? *Circulation*, 131(25):2167–2168, 2015.

[72] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[73] Barnabás Homola and Jakob E. Bardram. Research package – a flutter framework for mhealth research, 2019.

[74] Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, page 103801, 2020.

[75] Sheng Hu, Zhenzhou Shao, and Jindong Tan. A real-time cardiac arrhythmia classification system with wearable electrocardiogram. In *2011 International Conference on Body Sensor Networks*, pages 119–124. IEEE, 2011.

[76] Yu Hen Hu and Jeng-Neng Hwang. Handbook of neural network signal processing, 2002.

[77] Tero Hurnanen, Eero Lehtonen, Mojtaba Jafari Tadi, Tom Kuusela, Tuomas Kiviniemi, Antti Saraste, Tuija Vasankari, Juhani Airaksinen, Tero Koivisto, and Mikko Pänkäälä. Automated detection of atrial fibrillation based on time–frequency analysis of seismocardiograms. *IEEE journal of biomedical and health informatics*, 21(5):1233–1241, 2016.

[78] AliveCor Inc. *Available [Online]: https://www.alivecor.com/kardiamobile*, 2020.

[79] Apple Inc. Taking an ecg with the ecg app on apple watch series 4 or later, 2020.

[80] Marija D Ivanovic, Vladimir Atanasoski, Alexei Shvilkin, Ljupco Hadzievski, and Aleksandra Maluckov. Deep learning approach for highly specific atrial fibrillation and flutter detection based on rr intervals. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1780–1783. IEEE, 2019.

[81] Craig T January, L Samuel Wann, Joseph S Alpert, Hugh Calkins, Joaquin E Cigarroa, Joseph C Cleveland Jr, Jamie B Conti, Patrick T Ellinor, Michael D Ezekowitz, Michael E Field, et al. 2014 aha/acc/hrs guideline for the management of patients with atrial fibrillation: executive summary: a report of the american college of cardiology/american heart association task force on practice guidelines and the heart rhythm society. *Circulation*, 130(23):2071–2104, 2014.

[82] Søren Paaske Johnsen, Lene Worsaae Dalby, Tomas Täckström, Jens Olsen, and Anina Fraschke. Cost of illness of atrial fibrillation: a nationwide study of societal impact. *BMC health services research*, 17(1):714, 2017.

[83] R Burke Johnson and Anthony J Onwuegbuzie. Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7):14–26, 2004.

[84] William B Kannel. Coronary heart disease and atrial fibrillation: the framingham study. *Am heart J*, 106:389–396, 1983.

[85] Chakkalakkal Prabhakaran Karunadas and Cibu Mathew. Comparison of arrhythmia detection by conventional holter and a novel ambulatory ecg system using patch and android app, over 24 h period. *Indian Pacing and Electrophysiology Journal*, 20(2):49–53, 2020.

[86] Arnold M Katz. *Physiology of the Heart*. Lippincott Williams & Wilkins, 2010.

[87] Alan Kennedy, Dewar D Finlay, Daniel Guldenring, Raymond R Bond, Kieran Moran, and James McLaughlin. Automated detection of atrial fibrillation using rr intervals and multivariate-based classification. *Journal of electrocardiology*, 49(6):871–876, 2016.

[88] L Khadra, AS Al-Fahoum, and H Al-Nashash. Detection of life-threatening cardiac arrhythmias using the wavelet transformation. *Medical and Biological Engineering and Computing*, 35(6):626–632, 1997.

[89] Ahsan H Khandoker, Marimuthu Palaniswami, and Chandan K Karmakar. Support vector machines for automated recognition of obstructive sleep apnea syndrome from ecg recordings. *IEEE transactions on information technology in biomedicine*, 13(1):37–48, 2008.

[90] Hamid Khorrami and Majid Moavenian. A comparative study of dwt, cwt and dct transformations in ecg arrhythmias classification. *Expert systems with Applications*, 37(8):5751–5757, 2010.

[91] Michael H Kim, Stephen S Johnston, Bong-Chul Chu, Mehul R Dalal, and Kathy L Schulman. Estimation of total incremental health care costs in patients with atrial fibrillation in the united states. *Circulation: Cardiovascular Quality and Outcomes*, 4(3):313–320, 2011.

[92] S. Kiranyaz, T. Ince, and M. Gabbouj. Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, 2016.

[93] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, mar 2016.

[94] Paulus Kirchhof, Angelo Auricchio, Jeroen Bax, Harry Crijns, John Camm, Hans-Christoph Diener, Andreas Goette, Gerd Hindricks, Stefan Hohnloser, Lukas Kappenberger, et al. Outcome parameters for trials in atrial fibrillation: recommendations from a consensus conference organized by the german atrial fibrillation competence network and the european heart rhythm association. *Europace*, 9(11):1006–1023, 2007.

[95] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. How to evaluate technologies for health behavior change in hci research. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3063–3072. ACM, 2011.

[96] Matthieu Komorowski and Leo Anthony Celi. Will artificial intelligence contribute to overuse in healthcare? *Critical care medicine*, 45(5):912, 2017.

[97] Devender Kumar, Thomas Nilson, and Jakob E. Bardram. Carp movisens sampling package, 2020.

[98] Devender Kumar, Thomas Nilson, and Jakob E. Bardram. Movisens flutter plugin, 2020.

[99] Devender Kumar, Sadasivan Puthusserypady, and Jakob Eyvind Bardram. Cachet-cadb. *Available [Online]: https://doi.org/10.11583/DTU.14547264.v1*, May, 2021.

[100] Abishek Thekkeyil Kunnath, Dilraj Nadarajan, Manesh Mohan, and Maneesha V Ramesh. wicard: A context aware wearable wireless sensor for cardiac monitoring. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1097–1102. IEEE, 2013.

[101] Soonil Kwon, Joonki Hong, Eue-Keun Choi, Euijae Lee, David Earl Hostallero, Wan Ju Kang, Byunghwan Lee, Eui-Rim Jeong, Bon-Kwon Koo, Seil Oh, et al. Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study. *JMIR mHealth and uHealth*, 7(6):e12770, 2019.

[102] Pablo Laguna, Roger G Mark, A Goldberg, and George B Moody. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in cardiology 1997*, pages 673–676. IEEE, 1997.

[103] Olli Lahdenoja, Tero Hurnanen, Zuhair Iftikhar, Sami Nieminen, Timo Knuutila, Antti Saraste, Tuomas Kiviniemi, Tuija Vasankari, Juhani Airaksinen, Mikko Pänkäälä, et al. Atrial fibrillation detection via accelerometer and gyroscope of a smartphone. *IEEE Journal of Biomedical and Health Informatics*, 22(1):108–118, 2017.

[104] Dakun Lai, Yuxiang Bu, Ye Su, Xinshu Zhang, and Chang-Sheng Ma. Non-standardized patch-based ecg lead together with deep learning based algorithm for automatic screening of atrial fibrillation. *IEEE Journal of Biomedical and Health Informatics*, 24(6):1569–1578, 2020.

[105] Dakun Lai, Xinshu Zhang, Yifei Zhang, and Md Belal Bin Heyat. Convolutional neural network based detection of atrial fibrillation combing rr intervals and f-wave frequency spectrum. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4897–4900. IEEE, 2019.

[106] Rita Laureanti, Giulio Conte, Valentina DA Corino, Stefan Osswald, David Conen, Laurent Roten, Nicolas Rodondi, Peter Ammann, Christine S Meyer-Zuern, Leo Bonati, et al. Sex-related electrocardiographic differences in patients with different types of atrial fibrillation: Results from the swiss-af study. *International journal of cardiology*, 307:63–70, 2020.

[107] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, Jan 1997.

[108] Dingwen Li, Jay Vaidya, Michael Wang, Ben Bush, Chenyang Lu, Marin Kollef, and Thomas Bailey. Feasibility study of monitoring deterioration of outpatients using multimodal data collected by wearables. *ACM Transactions on Computing for Healthcare*, 1(1):1–22, 2020.

[109] Jin Peng Li, Damon Berry, and Richard Hayes. A mobile ecg monitoring system with context collection. In *4th European Conference of the International Federation for Medical and Biological Engineering*, pages 1222–1225. Springer, 2009.

[110] Leonard S Lilly. *Pathophysiology of heart disease: a collaborative project of medical students and faculty*. Lippincott Williams & Wilkins, 2012.

[111] Mohamed Limam and Frederic Precioso. Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[112] David T Linker. Accurate, automated detection of atrial fibrillation in ambulatory recordings. *Cardiovascular engineering and technology*, 7(2):182–189, 2016.

[113] Wendy E Mackay and Anne-Laure Fayard. Hci, natural science and design: a framework for triangulation across disciplines. In *Symposium on Designing Interactive Systems: Proceedings of the 2 nd conference on Designing interactive systems: processes, practices, methods, and techniques*, volume 18, pages 223–234, 1997.

[114] Athar N Malik and Brian V Nahed. Tremor mimicking ventricular tachycardia on telemetry. *World neurosurgery*, 147:158–159, 2021.

[115] Gabriela Marcu, Jakob E Bardram, and Silvia Gabrielli. A framework for overcoming challenges in designing persuasive monitoring and feedback systems for mental illness. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 1–8. IEEE, 2011.

[116] Manlio F Márquez, Luis Colín, Milton Guevara, Pedro Iturralde, and Antonio G Hermosillo. Common electrocardiographic artifacts mimicking arrhythmias in ambulatory monitoring. *American heart journal*, 144(2):187–197, 2002.

[117] Hannah Ramsden Marston, Robin Hadley, Duncan Banks, and María Del Carmen Miranda Duro. Mobile self-monitoring ecg devices to diagnose arrhythmia that coincide with palpitations: A scoping review. In *Healthcare*, volume 7, page 96. Multidisciplinary Digital Publishing Institute, 2019.

[118] Igor Matias, Nuno Garcia, Sandeep Pirbhulal, Virginie Felizardo, Nuno Pombo, Henriques Zacarias, Miguel Sousa, and Eftim Zdravevski. Prediction of atrial fibrillation using artificial intelligence on electrocardiograms: A systematic review. *Computer Science Review*, 39:100334, 2021.

[119] Anne Thushara Matthias and Jegarajah Indrakumar. Electrocardiogram artifact caused by rigors mimicking narrow complex tachycardia: a case report. *BMC research notes*, 7(1):1–3, 2014.

[120] David D McManus, Jo Woon Chong, Apurv Soni, Jane S Saczynski, Nada Esa, Craig Napolitano, Chad E Darling, Edward Boyer, Rochelle K Rosen, Kevin C Floyd, et al. Pulse-smart: pulse-based arrhythmia discrimination using a novel smartphone application. *Journal of cardiovascular electrophysiology*, 27(1):51–57, 2016.

[121] Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5, 2001.

[122] Wagner Meira, Antonio L. P. Ribeiro, Derick M. Oliveira, and Antonio H. Ribeiro. Contextualized interpretable machine learning for medical diagnosis. *Commun. ACM*, 63(11):56–58, October 2020.

[123] Fen Miao, Yayu Cheng, Yi He, Qingyun He, and Ye Li. A wearable context-aware ecg monitoring system integrated with built-in kinematic sensors of the smartphone. *Sensors*, 15(5):11465–11484, 2015.

[124] Suneet Mittal, Evgeny Pokushalov, Alexander Romanov, Martha Ferrara, Aysha Arshad, Dan Musat, Mark Preminger, Tina Sichrovsky, and Jonathan S. Steinberg. Long-term ecg monitoring using an implantable loop recorder for the detection of atrial fibrillation after cavotricuspid isthmus ablation in patients with atrial flutter. *Heart Rhythm*, 10(11):1598–1604, 2013.

[125] Miyasaka. Secular trends in incidence of atrial fibrillation in olmsted county, minnesota, 1980 to 2000, and implications on the projections for future prevalence (vol 114, pg 119, 2006). *Circulation*, 114(11):E498–E498, 2006.

[126] Jonathan Moeyersons, Elena Smets, John Morales, Amalia Villa, Walter De Raedt, Dries Testelmans, Bertien Buyse, Chris Van Hoof, Rik Willems, Sabine Van Huffel, and Carolina Varon. Artefact detection and quality assessment of ambulatory ecg signals. *Computer Methods and Programs in Biomedicine*, 182:105050, 2019.

[127] George Moody. A new method for detecting atrial fibrillation using rr intervals. *Computers in Cardiology*, pages 227–230, 1983.

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

[129] George B Moody, W Muldrow, and Roger G Mark. A noise stress test for arrhythmia detectors. *Computers in cardiology*, 11(3):381–384, 1984.

[130] Yun-Hong Noh, Jiunn Huei Yap, and Do-Un Jeong. Implementation of the abnormal ecg monitoring system using heartbeat check map technique. In *2013 International Conference on IT Convergence and Security (ICITCS)*, pages 1–4. IEEE, 2013.

[131] FM Nolle, FK Badura, JM Catlett, RW Bowser, and MH Sketch. Crei-gard, a new concept in computerized arrhythmia monitoring systems. *Computers in Cardiology*, 13:515–518, 1986.

[132] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[133] Godwin Okechukwu Ogbuabor, Juan Carlos Augusto, Ralph Moseley, and Aléchia van Wyk. Context-aware system for cardiac condition monitoring and management: a survey. *Behaviour & Information Technology*, pages 1–18, 2020.

[134] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[135] Julien Oster and Gari D Clifford. Impact of the presence of noise on rr interval-based atrial fibrillation detection. *Journal of electrocardiology*, 48(6):947–951, 2015.

[136] Saman Parvaneh, Jonathan Rubin, Saeed Babaeizadeh, and Minnan Xu-Wilson. Cardiac arrhythmia detection using deep learning: A review. *Journal of electrocardiology*, 57:S70–S74, 2019.

[137] Saman Parvaneh, Jonathan Rubin, Asif Rahman, Bryan Conroy, and Saeed Babaeizadeh. Analyzing single-lead short ecg recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation. *Physiological measurement*, 39(8):084003, 2018.

[138] Santosh Patel. Electrocardiographic artifact mimicking ventricular tachycardia during high-frequency oscillatory ventilation: A case report. *American Journal of Critical Care*, 15(3):310–311, 2006.

[139] Abdolrahman Peimankar and Sadasivan Puthusserypady. Dens-ecg: A deep learning approach for ecg signal delineation. *Expert Systems with Applications*, 165:113911, 2021.

[140] CR Lopez Perales, Harriette GC Van Spall, Shingo Maeda, Alejandro Jimenez, Decebal Gabriel Laţcu, Anat Milman, Fati Kirakoya-Samadoulougou, Mamas A Mamas, Daniele Muser, and Ruben Casado Arroyo. Mobile health applications for the detection of atrial fibrillation: a systematic review. *EP Europace*, 2020.

[141] Ming-Zher Poh, Yukkee Cheung Poh, Pak-Hei Chan, Chun-Ka Wong, Louise Pun, Wangie Wan-Chiu Leung, Yu-Fai Wong, Michelle Man-Ying Wong, Daniel Wai-Sing Chu, and Chung-Wah Siu. Diagnostic assessment of a deep learning

system for detecting atrial fibrillation in pulse waveforms. *Heart*, 104(23):1921–1928, 2018.

[142] Mark Pollet and Glenn N. Levine. Chapter 6 - ambulatory electrocardiogram monitoring. In Glenn N. Levine, editor, *Cardiology Secrets (Fifth Edition)*, pages 46–52. Elsevier, fifth edition edition, 2018.

[143] Bahareh Pourbabaee, Mehrsan Javan Roshtkhari, and Khashayar Khorasani. Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(12):2095–2104, 2018.

[144] Patrick X Quesnel, Adrian DC Chan, and Homer Yang. Signal quality and false myocardial ischemia alarms in ambulatory electrocardiograms. In *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5. IEEE, 2014.

[145] A. Rajkumar, M. Ganesan, and R. Lavanya. Arrhythmia classification on ecg using deep learning. In *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pages 365–369, 2019.

[146] Satish Ramkumar, Nitesh Nerlekar, Daniel D'Souza, Derek J Pol, Jonathan M Kalman, and Thomas H Marwick. Atrial fibrillation detection using single lead portable electrocardiographic monitoring: a systematic review and meta-analysis. *BMJ open*, 8(9):e024178, 2018.

[147] Meghan Reading, Dawon Baik, Melissa Beauchemin, Kathleen T Hickey, and Jacqueline A Merrill. Factors influencing sustained engagement with ecg self-monitoring: perspectives from patients and health care providers. *Applied clinical informatics*, 9(04):772–781, 2018.

[148] Iñaki Romero, Torfinn Berset, Dilpreet Buxi, Lindsay Brown, Julien Penders, Sunyoung Kim, Nick Van Helleputte, Hyejung Kim, Chris Van Hoof, and Firat Yazicioglu. Motion artifact reduction in ambulatory ecg monitoring: An integrated system approach. In *Proceedings of the 2nd Conference on Wireless Health*, WH '11, New York, NY, USA, 2011. Association for Computing Machinery.

[149] Jonathan Rubin, Saman Parvaneh, Asif Rahman, Bryan Conroy, and Saeed Babaeizadeh. Densely connected convolutional networks and signal quality analysis to detect atrial fibrillation using short single-lead ecg recordings. In *2017 Computing in cardiology (cinc)*, pages 1–4. IEEE, 2017.

[150] Norhashimah Mohd Saad, Abdul Rahim Abdullah, and Yin Fen Low. Detection of heart blocks in ecg signals by spectrum and time-frequency analysis. In *2006 4th Student Conference on Research and Development*, pages 61–65. IEEE, 2006.

[151] Saeed Saadatnejad, Mohammadhosein Oveisi, and Matin Hashemi. LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices. *IEEE Journal of Biomedical and Health Informatics*, 24(2):515–523, feb 2020.

[152] Milad Salem, Shayan Taheri, and Jiann-Shiun Yuan. Ecg arrhythmia classification using transfer learning from 2-dimensional deep cnn features. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2018.

[153] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. Hci and affective health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–17, New York, NY, USA, 2019. Association for Computing Machinery.

[154] Tommaso Sanna, Hans-Christoph Diener, Rod S. Passman, Vincenzo Di Lazzaro, Richard A. Bernstein, Carlos A. Morillo, Marilyn Mollman Rymer, Vincent Thijs, Tyson Rogers, Frank Beckers, Kate Lindborg, and Johannes Brachmann. Cryptogenic stroke and underlying atrial fibrillation. *New England Journal of Medicine*, 370(26):2478–2486, 2014. PMID: 24963567.

[155] Giovanna Sannino and Giuseppe De Pietro. A deep learning approach for ecg-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems*, 86:446–455, 2018.

[156] G Schreier, P Kastner, and W Marko. An automatic ecg processing algorithm to identify patients prone to paroxysmal atrial fibrillation. In *Computers in Cardiology 2001. Vol. 28 (Cat. No. 01CH37287)*, pages 133–135. IEEE, 2001.

[157] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[158] G Selvakumar, K Boopathy Bagan, and B Chidambararajan. Wavelet decomposition for detection and classification of critical ecg arrhythmias. In *Proceeding of the 8th WSEAS International Conference on Mathematics and computers in Biology and Chemistry*, pages 80–84. Citeseer, 2007.

[159] Mohamed Adel Serhani, Hadeel T. El Kassabi, Heba Ismail, and Alramzana Nujum Navaz. Ecg monitoring systems: Review, architecture, processes, and key challenges. *Sensors*, 20(6), 2020.

[160] Mohamed Adel Serhani, Hadeel T El Kassabi, Heba Ismail, and Alramzana Nujum Navaz. Ecg monitoring systems: Review, architecture, processes, and key challenges. *Sensors*, 20(6):1796, 2020.

[161] Supreeth Prajwal Shashikumar, Amit J Shah, Qiao Li, Gari D Clifford, and Shamim Nemati. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 141–144. IEEE, 2017.

[162] Qin Shen, Hongxiang Gao, Yuwen Li, Qi Sun, Minglong Chen, Jianqing Li, and Chengyu Liu. An open-access arrhythmia database of wearable electrocardiogram. *Journal of Medical and Biological Engineering*, 40(4):564–574, 2020.

[163] Yichen Shen, Maxime Voisin, Alireza Aliamiri, Anand Avati, Awni Hannun, and Andrew Ng. Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 1909–1916, New York, NY, USA, 2019. Association for Computing Machinery.

[164] Alireza Sahami Shirazi, Diana Cheng, Oliver Kroell, Dagmar Kern, and Albrecht Schmidt. Cardioviz: Contextual capture and visualization for long-term ecg data. In *adjunct Proceedings of Ubicomp*, 2007.

[165] Khairul A Sidek, Ibrahim Khalil, and Herbert F Jelinek. Ecg biometric with abnormal cardiac conditions in remote monitoring system. *IEEE Transactions on systems, man, and cybernetics: systems*, 44(11):1498–1509, 2014.

[166] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[167] Nikhil Singh, Sung Chun, David Hadley, and Victor Froelicher. Clinical implications of technological advances in screening for atrial fibrillation. *Progress in cardiovascular diseases*, 60(4-5):550–559, 2018.

[168] M Smoleń, E Kańtoch, P Augustyniak, and P Kowalski. Wearable patient home monitoring based on ecg and acc sensors. In *5th European Conference of the International Federation for Medical and Biological Engineering*, pages 941–944. Springer, 2011.

[169] Harold Smulyan. The computerized ecg: friend and foe. *The American journal of medicine*, 132(2):153–160, 2019.

[170] F Spadini, F Vergari, L Nachman, C Lamberti, and T Salmon Cinotti. A wireless and context-aware ecg monitor: An imote2 based portable system. In *2008 Computers in Cardiology*, pages 997–1000. IEEE, 2008.

[171] GS Stergiou, N Karpettas, A Protogerou, EG Nasothimiou, and M Kyriakidis. Diagnostic accuracy of a home blood pressure monitor to detect atrial fibrillation. *Journal of human hypertension*, 23(10):654–658, 2009.

[172] Emma Svennberg, Martin Stridh, Johan Engdahl, Faris Al-Khalili, Leif Friberg, Viveka Frykman, and Mårten Rosenqvist. Safe automatic one-lead electrocardiogram analysis in screening for atrial fibrillation. *EP Europace*, 19(9):1449–1453, 10 2016.

[173] Zenicor Medical Systems. The zenicor solution, 2020.

[174] A Taddei, G Distante, M Emdin, P Pisani, GB Moody, C Zeelenberg, and C Marchesi. The european st-t database: standard for evaluating systems for the analysis of st-t changes in ambulatory electrocardiography. *European heart journal*, 13(9):1164–1172, 1992.

[175] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[176] Anna Thorén, Araz Rawshani, Johan Herlitz, Johan Engdahl, Thomas Kahan, and Therese Djärv. Ecg monitoring in in-hospital cardiac arrest (ihca). *Resuscitation*, 142:e13, 2019. RESUSCITATION 2019 - Controversies in Resuscitation: Abstracts.

[177] Luan Tran, Yanfang Li, Luciano Nocera, Cyrus Shahabi, and Li Xiong. Multifusionnet: atrial fibrillation detection with deep neural networks. *AMIA Summits on Translational Science Proceedings*, 2020:654, 2020.

[178] HG Van Steenis and JHM Tulen. The effects of physical activities on cardiovascular variability in ambulatory situations [ecg/accelerometry analysis]. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.'Magnificent Milestones and Emerging Opportunities in Medical Engineering'(Cat. No. 97CH36136)*, volume 1, pages 105–108. IEEE, 1997.

[179] S Viskin, M Golovner, N Malov, R Fish, I Alroy, Y Vila, S Laniado, E Kaplinsky, and A Roth. Circadian variation of symptomatic paroxysmal atrial fibrillation. data from almost 10000 episodes. *European heart journal*, 20(19):1429–1434, 1999.

[180] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[181] Jibin Wang. A deep learning approach for atrial fibrillation signals classification based on convolutional and modified elman neural network. *Future Generation Computer Systems*, 102:670–679, 2020.

[182] John G Webster. *Medical instrumentation: application and design*. John Wiley & Sons, 2009.

[183] Xiaoling Wei, Jimin Li, Chenghao Zhang, Ming Liu, Peng Xiong, Xin Yuan, Yifei Li, Feng Lin, and Xiuling Liu. Atrial fibrillation detection by the combination of recurrence complex network and convolution neural network. *Journal of Probability and Statistics*, 2019, 2019.

[184] Peter West, Max Van Kleek, Richard Giordano, Mark J. Weal, and Nigel Shadbolt. Common Barriers to the Use of Patient-Generated Data Across Clinical Settings. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–13, 2018.

[185] Peter West, Max Van Kleek, Richard Giordano, Mark J Weal, and Nigel Shadbolt. Common barriers to the use of patient-generated data across clinical settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[186] Philip A Wolf, Robert D Abbott, and William B Kannel. Atrial fibrillation as an independent risk factor for stroke: the framingham study. *stroke*, 22(8):983–988, 1991.

[187] Meng-Hsi Wu and Edward Y Chang. Deepq arrhythmia database: A large-scale dataset for arrhythmia detector evaluation. In *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care*, pages 77–80, 2017.

[188] Ziqian Wu, Xujian Feng, and Cuiwei Yang. A deep learning method to detect atrial fibrillation based on continuous wavelet transform. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1908–1912. IEEE, 2019.

[189] Yong Xia, Naren Wulan, Kuanquan Wang, and Henggui Zhang. Detecting atrial fibrillation by deep convolutional neural networks. *Computers in biology and medicine*, 93:84–92, 2018.

[190] Zhaohan Xiong, Martyn P Nash, Elizabeth Cheng, Vadim V Fedorov, Martin K Stiles, and Jichao Zhao. Ecg signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network. *Physiological measurement*, 39(9):094006, 2018.

[191] Sean Shensheng Xu, Man-Wai Mak, and Chi-Chung Cheung. Towards end-to-end ecg classification with raw signal extraction and deep neural networks. *IEEE journal of biomedical and health informatics*, 23(4):1574–1584, 2018.

[192] Xiaoyan Xu, Shoushui Wei, Caiyun Ma, Kan Luo, Li Zhang, and Chengyu Liu. Atrial fibrillation beat identification using the combination of modified frequency slice wavelet transform and convolutional neural networks. *Journal of healthcare engineering*, 2018, 2018.

[193] Zhenjie Yao, Zhiyong Zhu, and Yixin Chen. Atrial fibrillation detection by multi-scale convolutional neural networks. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–6. IEEE, 2017.

[194] Özal Yıldırım, Paweł Pławiak, Ru-San Tan, and U Rajendra Acharya. Arrhythmia detection using deep convolutional neural network with long duration ecg signals. *Computers in biology and medicine*, 102:411–420, 2018.

[195] Yuming Hua, Junhai Guo, and Hua Zhao. Deep belief networks and deep learning. In *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things*, pages 1–4, 2015.

[196] Özgür Yürür, Chi Harold Liu, Zhengguo Sheng, Victor CM Leung, Wilfrido Moreno, and Kin K Leung. Context-awareness for mobile sensing: A survey and future directions. *IEEE Communications Surveys & Tutorials*, 18(1):68–93, 2014.

[197] I. Zagan, V. G. Gaitan, N. Iuga, and A. Brezulianu. m-greencardio embedded system designed for out-of-hospital cardiac patients. In *2018 International Conference on Development and Application Systems (DAS)*, pages 11–17, 2018.

[198] Wei Zhao, Jing Hu, Dongya Jia, Hongmei Wang, Zhenqi Li, Cong Yan, and Tianyuan You. Deep learning based patient-specific classification of arrhythmia on ecg signal. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1500–1503. IEEE, 2019.

[199] Qian Zheng, Chao Chen, Zhinan Li, Anpeng Huang, Bingli Jiao, Xiaohui Duan, and Linzhen Xie. A novel multi-resolution svm (mr-svm) algorithm to detect ecg signal anomaly in we-care project. In *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–6. IEEE, 2013.

[200] Xiaolin Zhou, Hongxia Ding, Benjamin Ung, Emma Pickwell-MacPherson, and Yuanting Zhang. Automatic online detection of atrial fibrillation based on symbolic dynamics and shannon entropy. *Biomedical engineering online*, 13(1):18, 2014.

[201] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. Convolutional recurrent neural networks for electrocardiogram classification. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[202] Peter Zimetbaum and Alena Goldman. Ambulatory arrhythmia monitoring: choosing the right device. *Circulation*, 122(16):1629–1636, 2010.

[203] W Zong, R Mukkamala, and RG Mark. A methodology for predicting paroxysmal atrial fibrillation based on ecg arrhythmia feature analysis. In *Computers in Cardiology 2001. Vol. 28 (Cat. No. 01CH37287)*, pages 125–128. IEEE, 2001.

[204] Massimo Zoni-Berisso, Fabrizio Lercari, Tiziana Carazza, and Stefano Domenicucci. Epidemiology of atrial fibrillation: European perspective. *Clinical epidemiology*, 6:213, 2014.