University of Cyprus

DEPARTMENT OF COMPUTER SCIENCE

Towards Privacy-Aware Usage of Fitness Trackers and Smart Home Devices: Enhancing User Awareness in the GDPR Era

Alexia Dini

A dissertation submitted to the University of Cyprus

in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

June, 2023

Alexia Dini

© Alexia Dini, 2023

Alexia Dini

VALIDATION PAGE

Doctoral Candidate: Alexia Dini

Doctoral Dissertation Title: Towards Privacy-Aware Usage of Fitness Trackers and Smart Home Devices: Enhancing User Awareness in the GDPR Era

*The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the Department of Computer Science and was approved on **June 9 , 2023** by the members of the Examination Committee.*

Examination Committee:

Research Supervisor
Associate Professor Georgia M. Kapitsaki

Committee Chair
Professor George A. Papadopoulos

Committee Member
Associate Professor Vasos Vassiliou

Committee Member
Associate Professor Michael Sirivianos

Committee Member
Associate Professor Andreas Constantinides

Alexia Dini

DECLARATION OF DOCTORAL CANDIDATE

The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Alexia Dini

. . . .

Alexia Dini

Abstract

The popularity of IoT devices has increased the collection, sharing and processing of user. The enormous amount of data collected and shared among these devices has raised a serious issue regarding the privacy and the awareness of the users about how the data generated by their devices are collected and shared. The EU General Data Protection Regulation aims to make the protection of personal data effective by providing a generic framework for the protection of the user and personal data privacy. However, in the era of smart homes and wearables, data transmitted to service providers may become available to interested third parties, who can process them with the intention to derive further knowledge and generate new insights and inferences about the users. Inferences have become one of the the biggest threats to privacy, compromising a basic privacy law, which is to allow a person to control who knows what about them. These issues put the user privacy at risk due to the inferences threat and the lack of user awareness about these inferences. Despite the research interest in the development of privacy-preserving methods to address the privacy challenges in IoT, the exploration of the users perspective and needs have not been adequately addressed in the effort to provide user-centric privacy-preserving solutions in IoT. This is the gap that this doctoral thesis aims to address, asking: How can we increase the smart home devices and fitness trackers users awareness over their data and privacy protection?. The thesis aims to comprehend users awareness over their data and privacy and develop mechanisms to evaluate and increase their awareness, and educate them about the privacy risks associated with the use of smart home devices and fitness trackers. It presents the characteristics that a user-centric, privacy-preserving and GDPR-compliant framework in IoT should incorporate and introduces a conceptual framework that demonstrates how the users can be provided with

Alexia Dini

the functionalities needed to be in control of their personal data created by IoT devices. The thesis identifies the inferences that can be extracted from smart home devices and fitness trackers data through the application of machine learning techniques, using real datasets that have been created through specific scenarios. The experimental results provide insights into the types of inferences that can be made from smart home and fitness tracker data, and highlight the importance of making the users aware about them. Consequently, we contribute with a tool, PrivacyEnhAction, that aims to increase the user awareness about potential privacy vulnerabilities that emerge from the use of these devices. A qualitative user study was conducted to evaluate the impact of PrivacyEnhAction to the awareness of the participants regarding possible inferences from fitness tracker data, with positive results. To further assist the effort of increasing the awareness of the users, this doctoral thesis provides a methodology for the analysis of the text of fitness trackers and smart home devices privacy policies, which is also implemented in the PrivacyEnhAction web application.

Alexia Dini

Περίληψη

Η δημοτικότητα των συσκευών Διαδικτύου των Πραγμάτων, όπως οι έξυπνες οικιακές συσκευές και οι συσκευές παρακολούθησης της φυσικής κατάστασης (fitness trackers), έχει προάγει την απόκτηση, την ανταλλαγή και τη διανομή δεδομένων που δημιουργούνται. Ο τεράστιος όγκος δεδομένων που συλλέγεται και μοιράζεται εγείρει ένα σοβαρό ζήτημα σχετικά με το απόρρητο και την επίγνωση των χρηστών για τον τρόπο συλλογής και κοινής χρήσης των δεδομένων που δημιουργούνται από τις συσκευές τους. Ο Γενικός Κανονισμός για την Προστασία Δεδομένων στοχεύει στο να καταστήσει αποτελεσματική την προστασία των προσωπικών δεδομένων, παρέχοντας ένα γενικό πλαίσιο για την προστασία του απορρήτου του χρήστη και των προσωπικών δεδομένων. Ωστόσο, στην εποχή του έξυπνου σπιτιού και των fitness trackers, τα δεδομένα που διαβιβάζονται σε παρόχους υπηρεσιών ενδέχεται να καταστούν διαθέσιμα σε ενδιαφερόμενα τρίτα μέρη, τα οποία μπορούν να τα επεξεργαστούν με σκοπό να εξάγουν συμπεράσματα για τους χρήστες. Η εξαγωγή συμπερασμάτων αποτελεί μια μεγάλη απειλή για το απόρρητο. Παρά το ερευνητικό ενδιαφέρον που υπάρχει για την ανάπτυξη μεθόδων προστασίας της ιδιωτικότητας για την αντιμετώπιση των προκλήσεων απορρήτου στο Διαδίκτυο των Πραγμάτων, η διερεύνηση της προοπτικής του χρήστη δεν έχει αντιμετωπιστεί στο Διαδίκτυο των Πραγμάτων. Αυτό είναι το ερευνητικό κενό που στοχεύει να αντιμετωπίσει αυτή η διατριβή, ρωτώντας: «Πώς μπορούμε να αυξήσουμε την επίγνωση των χρηστών των έξυπνων οικιακών συσκευών και των fitness trackers σχετικά με τα δεδομένα και την προστασία του απορρήτου τους». Η διατριβή στοχεύει να κατανοήσει την επίγνωση των χρηστών σχετικά με τα δεδομένα και το απόρρητό τους και να αναπτύξει μηχανισμούς αξιολόγησης και αύξησης της επίγνωσής τους, να τους εκπαιδεύσει σχετικά με τους κινδύνους απορρήτου που σχετίζονται με τη χρήση έξυπνων οικιακών συσκευών και fitness trackers και να τους ενδυναμώσει με τον έλεγχο των δεδομένων και του απορρήτου τους στο πλαίσιο αυτό. Παρουσιάζονται τα χαρακτηριστικά που πρέπει να διαθέτει ένα συμβατό με το GDPR πλαίσιο με επίκεντρο τον χρήστη και την διατήρηση της ιδιωτικότητας, και ένα

Alexia Dini

πλαίσιο που βασίζεται σε αυτά τα χαρακτηριστικά που δείχνει πώς μπορούν να παρέχονται στους χρήστες οι λειτουργίες που χρειάζονται, για να έχουν τον έλεγχο των προσωπικών τους δεδομένων. Η διατριβή προσδιορίζει τα πιθανά συμπεράσματα που μπορούν να εξαχθούν από δεδομένα έξυπνων οικιακών συσκευών και fitness trackers μέσω της εφαρμογής τεχνικών μηχανικής μάθησης, χρησιμοποιώντας πραγματικά σύνολα δεδομένων που έχουν δημιουργηθεί μέσω συγκεκριμένων σεναρίων. Τα πειραματικά αποτελέσματα παρέχουν πληροφορίες για τους τύπους συμπερασμάτων που μπορούν να εξαχθούν και υπογραμμίζουν τη σημασία της ενημέρωσης του χρήστη σχετικά με αυτά. Στην διατριβή παρουσιάζεται το εργαλείο «PrivacyEnhAction» που στοχεύει στην αύξηση της επίγνωσης των χρηστών σχετικά με πιθανές ευπάθειες απορρήτου που προκύπτουν από τη χρήση αυτών των συσκευών. Παράλληλα, έχει διεξαχθεί μια ποιοτική μελέτη χρηστών για να αξιολογηθεί ο αντίκτυπος του PrivacyEnhAction στην επίγνωση των συμμετεχόντων σχετικά με τα πιθανά συμπεράσματα που μπορούν να εξαχθούν από τα δεδομένα από fitness trackers, με θετικά αποτελέσματα. Για να βοηθήσει περαιτέρω στην προσπάθεια αύξησης της επίγνωσης των χρηστών, η διατριβή παρέχει μια μεθοδολογία για την ανάλυση του κειμένου της Πολιτικής Απορρήτου έξυπνων οικιακών συσκευών και fitness trackers, η οποία έχει υλοποιηθεί και στο εργαλείο PrivacyEnhAction.

Alexia Dini

Acknowledgments

First and foremost, I am incredibly grateful to my supervisor, Assoc. Prof. Georgia M. Kapitsaki, for her guidance, invaluable advice, continuous support and patience during my PhD study. Her deep knowledge and experience have guided me at all times during my academic research. Working with them during this time has been an honour and a pleasure. Thank you, Georgia, for endless discussions, ideas, and your commitment to developing me as a researcher. Also, I would like to express my sincere gratitude and appreciation to the members of the committee, Prof. George Papadopoulos and Assoc. Prof. Vassos Vassiliou, for their invaluable contribution to my PhD journey. Their expertise, guidance and constructive feedback have been instrumental in the successful completion of my doctoral thesis. I am also most grateful to Assoc. Prof. Ioannis Katakis, for providing valuable feedback, ideas, and insightful discussions throughout my research.

None of this would have been possible without the incredible support of my family. To my husband Tasos, thank you for your patience, understanding, and unwavering support. Your love and encouragement have been my greatest source of strength and motivation during this challenging journey. To my children, Anastasia and Vasiliki, the sunshine of my life, thank you for bringing joy, laughter, and inspiration to my life. Your love and support have been the driving force behind my determination to succeed. I could not have achieved this without the love and support of my family, and for that, I am forever grateful.

Alexia Dini

Publications

Journal Articles

1. **A. Dini Kounoudes**, G.M. Kapitsaki, I. Katakis, "Enhancing user awareness on inferences obtained from fitness trackers data", *User Modeling and User-Adapted Interaction*, pp. 1-48, Jan 2023.

2. **A. Dini Kounoudes**, G.M. Kapitsaki, "A mapping of IoT user-centric privacy preserving approaches to the GDPR", *Internet of Things*, 11, p. 100179, Sep 2020.

Conference Proceedings

3. **A. Dini Kounoudes**, G.M. Kapitsaki, I. Katakis, M. Milis. "User-centred privacy inference detection for smart home devices", in *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation*, Atlanta, USA, 2021, pp. 210-218.

4. G.M. Kapitsaki, **A. Dini Kounoudes**, A.P. Achilleos. "An overview of user privacy preferences modeling and adoption", in *46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Portoro, Slovenia, 2020, pp. 569-576.

5. **A. Dini Kounoudes**, G.M. Kapitsaki, M. Milis. "Towards considering user privacy preferences in smart water management", in *27th Conference on User Modeling, Adaptation and Personalization*, Larnaca, Cyprus, 2019, pp. 209-212.

Conference Posters

6. **A. Dini Kounoudes**. "The open area of user-centric privacy protection in the Internet of Things", in *6th ACM Celebration of Women in Computing: womENcourage 2019*, Rome, Italy, 2019.

Alexia Dini

Contents

1 Introduction 1

1.1 Motivation 2

1.2 Research Questions 4

1.3 Research Contributions 7

1.4 Thesis Outline 8

2 Background and Related Work 13

2.1 Smart Home and Wearable Devices 13

2.2 Privacy Definitions 15

2.3 The General Data Protection Regulation 16

2.4 Privacy protection in IoT 19

2.4.1 Privacy protection frameworks 19

2.4.2 Privacy protection in the Smart Home 21

2.4.3 Privacy protection in Wearables 22

2.5 Privacy Inference Risks in IoT 24

2.5.1 Data collection in IoT 24

2.5.2 Information inference as a privacy threat in IoT 25

2.5.3 Inferences from Smart Home devices 26

2.5.4 Inferences from fitness trackers data 28

2.6 User awareness and concerns on IoT privacy 32

2.6.1 User awareness and privacy concerns of smart home devices users . 33

2.6.2 User awareness and privacy concerns of fitness trackers users . . . 35

2.6.3 Information inference threats in other domains 37

2.7 Privacy policy analysis as a tool for user privacy awareness 37

2.8 Machine Learning Techniques 40

2.8.1 Supervised Machine Learning Algorithms 40

2.8.2 Unsupervised Machine Learning Algorithms 45

2.9 Discussion and Conclusion 48

Alexia Dini

3 **A user-centric privacy framework for personal data protection in IoT** **51**
3.1 Background . . . 52
3.2 The GDPR Challenges . . . 53
3.3 Characteristics and mapping to challenges . . . 56
3.4 Challenge 1: Profiling, inference and discrimination . . . 60
3.4.1 CR1 - Prevent inference . . . 60
3.4.2 CR2 - Provide data transformation . . . 61
3.5 Challenge 2: Control and context-sensitive sharing . . . 63
3.5.1 CR4 - Provide control of personal data to users . . . 63
3.5.2 CR6 - Provide tools for data management to users . . . 64
3.5.3 CR5 - Provide monitoring and control of devices that collect data . 64
3.5.4 CR7 - Provide ability for data erasure . . . 65
3.5.5 CR8 - Provide transparency . . . 65
3.5.6 CR9 - Provide balance of privacy between user and third parties . . 66
3.6 Challenge 3: Consent and uncertainty . . . 67
3.6.1 CR2 - Provide data transformation . . . 67
3.6.2 CR4 - Provide control of personal data to users . . . 68
3.6.3 CR6 - Provide tools for data management to users . . . 69
3.6.4 CR10 - Provide enforcement of user privacy preferences . . . 69
3.6.5 CR11 - Provide privacy by design or privacy by default . . . 73
3.6.6 CR12 - Provide ability to users to make informed consent choices . 74
3.6.7 CR13 - Estimate privacy risks of data collection/inference to users . 75
3.6.8 CR14 - Communicate risks of data collection/inference to users . . 75
3.6.9 CR15 - Provide ability to users to specify their privacy preferences . 76
3.6.10 CR16 - Prevent excessive data collection . . . 76
3.7 Challenge 4: Honesty, trust, and transparency . . . 77
3.7.1 CR3 - Provide user awareness on data collection . . . 77
3.7.2 CR8 - Provide transparency . . . 78
3.8 The proposed user-centric privacy protection framework for IoT . . . 78
3.8.1 Description of Privacy-EnhAction steps . . . 79
3.9 Discussion . . . 83

4 **Inference detection in a smart home scenario** **89**
4.1 The Smart Home scenario . . . 90

Alexia Dini

4.2 Methodology . . . 91

4.2.1 Overview . . . 91

4.2.2 Data processing and cleaning . . . 93

4.2.3 Cluster analysis . . . 95

4.2.4 Interpretation of Clusters . . . 97

4.2.5 Inferences detection . . . 98

4.3 The Smart Home Experiments . . . 98

4.3.1 Data set description . . . 98

4.3.2 Exploratory analysis . . . 98

4.4 Experimental results . . . 99

4.4.1 Comparison of clustering algorithms . . . 99

4.4.2 Rules definition and cluster interpretation . . . 102

4.4.3 Extracting Inferences . . . 104

4.5 Proof of concept application . . . 106

4.6 Discussion . . . 107

5 Inference detection in the fitness trackers domain 108

5.1 Possible inferences from fitness trackers data . . . 108

5.2 Fitness trackers scenarios under study . . . 109

5.2.1 Data collection process . . . 109

5.2.2 Data processing and cleaning . . . 112

5.2.3 Data analysis techniques . . . 114

5.3 Inference identification in fitness trackers under study . . . 114

5.3.1 Fitbit inference detection analysis . . . 114

5.3.2 Garmin inference detection analysis . . . 116

5.3.3 Mi Fit inference detection analysis . . . 117

5.4 Implementation . . . 118

5.4.1 Extensions to the PrivacyEnhAction tool . . . 119

5.5 Discussion . . . 121

6 Evaluating the impact of PrivacyEnhAction to users awareness 123

6.1 Research Questions . . . 123

6.2 Material and methods: Empirical approach . . . 124

6.2.1 Research participants recruiting . . . 125

Alexia Dini

6.3 Questionnaire on fitness trackers user privacy concerns 125
6.3.1 Demographics and other results 126
6.3.2 User awareness on privacy risks 127
6.3.3 User privacy concerns 130
6.4 PrivacyEnhAction Evaluation Questionnaire 131
6.4.1 PrivacyEnhAction usability results 131
6.4.2 User awareness on privacy risks 131
6.4.3 User opinions about PrivacyEnhAction 134
6.5 Discussion 135

7 Analysis of smart devices privacy policies 139
7.1 Fitness trackers privacy policies 140
7.1.1 FitBit privacy policy regarding data sharing 140
7.1.2 Xiaomi privacy policy regarding data sharing 143
7.2 Smart home devices privacy policies 144
7.2.1 Vivint privacy policy regarding data sharing 144
7.2.2 Arlo privacy policy regarding data sharing 146
7.2.3 Hive privacy policy regarding data sharing 146
7.2.4 Philips Hue privacy policy regarding data sharing 147
7.3 Discussion 147

8 Extracting GDPR user rights and inference risks from privacy policy texts 149
8.1 Contributions 149
8.2 Approach overview 150
8.3 Methodology Steps 152
8.3.1 Privacy policies text collection 152
8.3.2 Definition of GDPR user rights terms list 153
8.3.3 Definition of inference groups 154
8.3.4 Privacy policy corpora preparation 160
8.3.5 Data pre-processing 162
8.3.6 SpotAware multi-label classification 163
8.3.7 Multi-label classification evaluation 164
8.4 Experimental Setup 165
8.4.1 SpotAware overall results 166

Alexia Dini

8.4.2 SpotAware results per user right and inference group 168

8.4.3 Results summarization 170

8.5 Extending PrivacyEnhAction with the SpotAware approach 172

8.6 Discussion 173

9 The implementation of the PrivacyEnhAction application 178

9.1 Introduction 178

9.2 PrivacyEnhAction Design and Architecture 179

9.2.1 Overview of the PrivacyEnhAction architecture 179

9.2.2 Overview of the PrivacyEnhAction design 179

9.3 PrivacyEnhaction Implementation 181

9.3.1 Back-end development of PrivacyEnhAction 181

9.3.2 Front-end development of PrivacyEnhAction 182

9.4 PrivacyEnhAction Features and Modules 182

9.4.1 Inference Detection Analysis module 182

9.4.2 Privacy Policy Analysis module 188

9.5 Further Work 190

9.6 Conclusion 191

10 Conclusions and Future Work 193

10.1 Introduction 193

10.2 Review of Research Goals and Findings 194

10.2.1 What are the characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess? 194

10.2.2 What inferences can be made from data collected from smart home devices and fitness trackers? 196

10.2.3 Are the users aware of the inferences that can be made about them from their fitness trackers data? 198

10.2.4 Can we enhance the awareness of the users regarding the possible inferences that can be obtained from their fitness tracker data? . . . 199

10.3 Thesis Contributions 199

10.4 Open Research Issues 203

10.5 Takeaway messages 205

10.6 Conclusions and Future Work 206

Alexia Dini

A List of GDPR user right Terms 233

A.1 The Right to Information Terms File 233

A.2 The Right of Access Terms file 235

A.3 The Right to Rectification Terms File 236

A.4 The Right to Erasure Terms File 236

A.5 The Right to Restriction of Processing Terms File 237

A.6 The Right to Data Portability Terms File 237

A.7 The Right to Object Terms File 238

A.8 The Right to Avoid Automated Decision-Making Terms File 238

Alexia Dini

Abbreviations

AAL	Ambient Assisted Living
AI	Artificial Intelligence
ASM	Advanced Sleep Monitoring
BLE	Bluetooth Low Energy
BR	BinaryRelevance
CC	ClassifierChains
DBSCAN	Density-based spatial clustering of applications with noise
DES	Data Encryption Standard
EDA	Exploratory Data Analysis
GDPR	General Data Protection Regulation
GPS	Global Positioning System
H-IoT	Health-related Internet of Things
IDC	Inference Detection Component
IoToys	Internet of Toys
IoT	Internet of Things
JSON	JavaScript Object Notation
kNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
LSB	Least Significant Bit
ML	Machine Learning
NAZ	No Activity Zones

Alexia Dini

NFC	Near Field Communications
NLP	Natural Language Processing
NLT	Natural Language Toolkit
PbD	Privacy by Design
PII	Personally Identifiable Information
PIR	Passive Infrared Sensor
RCC	Rights Classification Component
SVM	Support Vector Machine
TF-IDF	Term-Frequency Inverse Document frequency
WSGI	Web Server Gateway Interface

Alexia Dini

List of Figures

2.1 Agglomerative Hierarchical Clustering procedure [148] 48

2.2 The flowchart of the Mean-shift clustering algorithm [182] 48

3.1 Proposed architecture of the proposed user-centric privacy framework, PrivacyEnhAction 79

3.2 Proposed steps for the development of a user-centric privacy framework for IoT 80

4.1 A 3-D image of the smart water meter 90

4.2 Possible inferences from smart water meter and motion sensor data 92

4.3 Example from the motion sensors dataset 94

4.4 Results of the K-means Algorithm for the Smart Water Meter data using the NAZ approach 102

4.5 Results of the DBSCAN Algorithm for the Smart Water Meter data using the NAZ approach 103

4.6 Results of the Agglomerative Hierarchical Clustering Algorithm for the Smart Water Meter data using the NAZ approach 103

4.7 List of rules derived in the smart home scenario experiments 104

4.8 Initial PrivacyEnhAction Inference Detection page 106

4.9 PrivacyEnhAction results for Smart Motion sensor data 107

5.1 PrivacyEnhAction Inference Detection Analysis page 119

5.2 Screenshot from PrivacyEnhAction Fitbit data inferences results page . . . 120

5.3 PrivacyEnhAction: Fitbit inferences from heart rate data 121

5.4 Taxonomy of Inferences from Smart Home Devices and Fitness Trackers . 122

6.1 Users attitudes with regards to privacy policies 127

6.2 User awareness on fitness tracker data collection and sharing 128

6.3 Normality test and descriptive statistics on Questionnaire 1 Section 5 data . 129

6.4 Users responses on Questionnaire 1 Section 5 129

6.5 Examples of user privacy concerns regarding the use of fitness trackers . . . 130

6.6 User awareness regarding the data collected by fitness trackers 131

Alexia Dini

6.7 Examples of users possible actions after using PrivacyEnhAction 134

8.1 Overview of our approach 151
8.2 PrivacyEnhAction Privacy Policy Analysis page 171
8.3 Results of the GDPR User Rights Analysis component from PrivacyEnhAction 172
8.4 Results of the Inferences Analysis component from PrivacyEnhAction . . . 173
8.5 Average score of GDPR compliance per sector 175
8.6 Average number of GDPR terms per user right per sector 175
8.7 Frequency of occurrence of inference groups per privacy policy per sector . 176

9.1 PrivacyEnhAction back-end architecture [329] 179
9.2 PrivacyEnhAction Inference Detection Analysis page 184
9.3 Garmin Inference Detection Analysis results 184
9.4 Garmin Activity and Location Inferences 185
9.5 Garmin Fitness Inferences 186
9.6 Garmin Sleep Inferences 187
9.7 PrivacyEnhAction Privacy Policy Analysis page 188
9.8 Results about GDPR User Rights Compliance for Withings 189
9.9 Results about Inference Detection for Withings 190

Alexia Dini

List of Tables

2.1 Summary of related work for privacy protection in IoT 20

2.2 Summary of related work for privacy inference risks in IoT 23

2.3 Activity levels and steps indices [292] 30

2.4 Summary of related work for user awareness and concerns on IoT privacy . 34

2.5 Summary of related work for privacy policy analysis as a tool for user privacy awareness 38

3.1 Mapping of GDPR characteristics to challenges 60

3.2 Approaches for Profiling, Inference and Discrimination - CH1 63

3.3 Approaches for Control and Context-sensitive sharing - CH2 66

3.4 Approaches for Consent and Uncertainty - CH3 70

3.5 Approaches for Honesty, Trust and Transparency - CH4 77

3.6 Proposed steps in the Privacy-EnhAction framework 82

3.7 Challenges and techniques used in literature related to the characteristics . . 84

4.1 Results for Smart Water Meter data using the consumption-based approach 105

5.1 Possible fitness trackers inferences 110

5.2 Indicative list of available commercial fitness trackers 111

5.3 Participants demographics 112

5.4 Fitness tracker datasets downloaded from repositories 112

6.1 Information about fitness tracker brands being used by the survey participants 126

6.2 Information about length of time of owning a fitness tracker 127

6.3 Overall mean scores on user awareness before and after interacting with the PrivacyEnhAction application 132

6.4 Paired-samples t-test results 133

7.1 Comparison of FitBit, Garmin and Xiaomi fitness trackers privacy policies . 141

7.2 Comparison of Vivint, Arlo, Hive and Philips Hue privacy policies 145

8.1 Number of terms in each GDPR user right list created for SpotAware 152

8.2 Example of GDPR User Rights Terms List 154

Alexia Dini

8.3 List of inference groups deducted from privacy policies 154
8.4 Inferences found in literature 157
8.5 Statistics of privacy policy corpora 162
8.6 SpotAware Rights Classification component - Classification results 166
8.7 SpotAware Inference Detection component - Classification results 167
8.8 RCC - BR and CC prediction performance for combined dataset 169
8.9 IDC - BR and CC prediction performance for combined dataset 170

Alexia Dini

Chapter 1

Introduction

Alexia Dini

With the ability to connect and control billions of devices and get access to valuable data, the Internet of Things (IoT) is shaping the future of technology and society, as it is estimated that the number of connected devices will rise to 50 billion by 2030 [298]. We can already see how the IoT has transformed the world in various domains throughout the last decade, by providing faster transportation systems, safer street lighting and energy-efficient buildings in smart cities, or by helping doctors to have faster access to patients data, to accurately track vital patient diagnostics, like heart rate or blood pressure, or to provide assistance to the elderly by allowing caretakers to monitor them on a 24 hour basis.

As the use of IoT devices is rising, sensor technology becomes smaller, more powerful and inexpensive, while data streams are becoming more accessible. Today, IoT devices have become very popular creating an explosion on the data generated and shared. The amount of data shared between IoT devices is prodigious, as "*it is estimated that today the average person creates 1.5 GB of data on average daily*" [168]. It is no wonder that the phrase "*data is the new gold*" [48, 102, 103] is a metaphor describing a new paradigm that revolutionises the world nowadays [83].

The enormous amount of data collected and shared among IoT resources has raised a serious issue regarding the privacy and the awareness of the users about how the data generated by their IoT connected devices are collected and shared. In a interview in 2000, the late Andrew Grove, CEO of Intel Corporation, prophetically stated that "*privacy is one of the biggest problems in this new electronic age*" [81]. The protection of personal data forms a principal citizen right safeguarded in the European Union that is particularly pertinent in the IoT domain. The EU General Data Protection Regulation (GDPR) [93], introduced in 2018, aims to make the protection of this right effective by providing a high level of data protection, and fundamental directions to accomplish an equitable treatment of the third parties or services and the users. Additionally, it radically changes how data are handled in every area applied, from industry to energy and others, and creates standards for the protection of user data in IoT, such as informed consent or privacy by design, among others. Furthermore, the regulation intends to provide a generic framework for the protection of the user and personal data privacy and to provide awareness to the users of how their data are collected and processed.

1.1 Motivation

Privacy has been recognised as a fundamental human right by the United Nations member states since 1948 [205]. The GDPR strengthens the user rights, introduces new standards for the handling of data, and requires the enforcement of privacy by design by the relevant technologies. Since the IoT depends upon effusive user data collection and sharing in order for the user to be uniquely involved in the process, this attribute increases the probability of risks for the user privacy. The GDPR provides a number of ways to address these risks, which "*however may be insufficient to ensure a fair balance between users' and providers' interests*" [307], while it calls for "*increased user involvement in protecting their data by enabling them to control what is collected about them, when, by whom and for what purposes*" [26]. In this respect, traditional privacy approaches must advance with moving their focus from the service providers to the users, giving them the power to control the exposure of their personal data. Studies of the recent literature regarding the protection of the user privacy in IoT environments have shown that work in this area is still at early stages, with research focusing mostly on technology and legal solutions [276]. In order for the users to be included in the data protection process in a beneficial way, they must be notified about and comprehend the privacy risks related to the exposure of their data to third parties. Furthermore, they must be able to counterpoise those risks with any possible advantages they will gain by making their data available to services or third parties. To that end, it is important that users are able to take "*meaningful privacy decisions regarding whether to disclose their data or not and to which extent*" [26].

In the IoT era, sensitive and non-sensitive data are recorded and transmitted to multiple service providers and platforms, aiming to improve the quality of our lives through the provision of high-quality services. However, in some cases these data may become available to interested third parties, who can analyse them with the intention to derive further knowledge and generate new insights and inferences about the users, that they can ultimately use

Alexia Dini

for their own benefit. Inferences have become one of the the biggest threats to privacy, due to the development of sophisticated machine learning techniques [309] and big data analytics, that are being used for the extraction of useful inferences from apparently harmless data or identified behaviour, compromising a basic privacy law, which is to allow a person to control who knows what about them [134]. Similar techniques are also used for making predictions about people's private lives, behaviours, habits and preferences, establishing the perfect conditions for discrimination, prejudicial and intrusive decision-making against the people involved [309]. This predicament raises a crucial issue regarding the privacy of the users and their awareness on how their personal data created by their IoT devices are shared and potentially used.

The research community has been actively working on developing privacy-preserving methods to address the privacy challenges in IoT, however in the existing literature, limited attention has been given to the development of user awareness mechanisms that can assist the users in understanding how the data created by their smart devices can be exploited for the extraction of inferences regarding their daily activities and lifestyle in general. There is an imperative need for the development of such tools, as IoT devices collect sensitive personal information that can be acquired by unauthorised third parties without user awareness [160], and also because these devices have become the perfect prey for attacks and data breaches, due to the lack of strict security guidelines and the sensitive nature of the data collected by them [191]. At present, existing awareness mechanisms come in the form of lengthy privacy policies [12] that the users generally tend to ignore; thus, further research is required in order to design the necessary tools and approaches to make the users aware of how their smart devices data can be exploited by third parties presenting the information in a direct and comprehensive way [162] and enable them to assimilate how to reduce these risks, by suggesting simple solutions as for example by altering their privacy preferences.

Smart home devices and fitness trackers are among the most popular IoT devices. Smart home devices are increasingly being adopted in households, and they include products such as smart speakers, smart thermostats, smart lighting, smart locks, and security cameras. These devices offer convenience and enhance the quality of life for their users by allowing them to control various aspects of their home environment remotely. According to a report by Statista, the global smart home market is expected to reach over $222.9 billion by 2027 [270]. Fitness trackers, on the other hand, are popular wearable devices that monitor and track various aspects of a person's physical activity and health. These devices can track the number of steps taken, calories burned, heart rate, and sleep patterns. Fitness trackers

Alexia Dini

have become increasingly popular over the years, with some of the most popular brands being Fitbit, Garmin, and Apple Watch. According to a report by Grand View Research, the global wearable device market, including fitness trackers, is expected to reach $186.14 billion by 2030 [121]. Given the widespread adoption of smart home devices and fitness trackers and the increasing amount of personal data that they collect, in this doctoral thesis we have selected them as our focus areas, aiming to address the issues we have identified under these two domains and propose potential solutions to address them.

The remainder of this Chapter explores the thesis from multiple dimensions, including the research questions, the contributions, and the thesis organisation.

1.2 Research Questions

In this context, the research problem that we address in this doctoral thesis is:

How can we increase the smart home devices and fitness trackers users' awareness over their data privacy protection?

Through our research, we approached our research problem through the following research questions that unfolded in the progress:

RQ1 - What are the characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess?.

As user privacy awareness is a critical issue in IoT, the GDPR addresses this concern by requiring IoT manufacturers and service providers to implement the appropriate measures in order to increase the awareness of the users of their rights regarding the protection of their privacy. Lack of transparency is a challenge that the users of smart devices have to face, as IoT service providers do not provide simple and concise information about their data practices in relevance to the data they collect, how they use these data or who they share these data with, making it challenging for the users to understand how their data are being processed and make informed decisions about their privacy. Furthermore, the lack of the provision of sufficient information about data collection and processing practices, enables IoT manufacturers or service providers to obtain user consent, with the users agreeing to data processing without fully understanding the risks associated with it. The large amounts of user data collected and shared by their smart devices and the limited control of the users over these data, also raise an important issue regarding the privacy and the awareness of the users about how their data are collected and shared and constitute another challenge that we have

Alexia Dini

identified in our research. In order to address these issues, a user-centric privacy framework is essential for protecting personal data in the IoT, which should provide users with greater control over their personal data and increase their awareness of their data and privacy protection. The objective of this question is to define such a framework and the characteristics that it should possess to ensure compliance with GDPR regulations and protect user privacy.

RQ2 - What inferences can be made from data collected from smart home devices and fitness trackers?.

In this thesis we have identified inference risks as a challenge in the protection of user data and privacy in IoT. Inference is defined as “*a conclusion that you draw about something by using information that you already have about it*" [63]. Inference risks can be defined as the potential for personal information to be inferred or derived from seemingly innocuous or unrelated data points. In IoT, this can occur when data from multiple sources are combined and analysed to draw conclusions about an individual’s behaviour, preferences, or other personal characteristics. Inference risks can present a significant challenge in the protection of user data and privacy in IoT as they raise critical concerns related to the privacy of the users and their awareness of how their personal data created by their IoT devices are shared and potentially used. For example, data from a smart home device such as a thermostat may be combined with data from a fitness tracker to infer whether an individual is at home or not. This information can then be used by advertisers or other third parties to make decisions about the individual’s eligibility for certain services. The challenge with inference risks is that they may not be immediately apparent to users, and they may not be able to control the information that is inferred from their data. This highlights the importance of transparency and user control in any privacy framework for IoT. Users should be informed about what data are being collected and how they may be combined or analysed, and they should have the ability to control how their data are used and shared. As in this thesis we focus on smart home devices and fitness trackers, the objective of this question is to identify the inferences that can be extracted from data collected from these devices.

RQ3 - Are the users aware of the inferences that can be made about them from their fitness trackers data?.

Research has shown that users may not be fully aware of the inferences that can be made about them from their IoT devices data. While some users may understand that their IoT devices collect data about their activities and behaviours, they may not fully comprehend

Alexia Dini

the extent to which these data can be used to make inferences about their personal lives. The objective of this question is to determine the level of awareness of fitness trackers users about the inferences that can be made from the data collected by their devices. The primary objective of the study was to investigate privacy concerns and risks associated with fitness trackers specifically. By narrowing the scope to fitness trackers, we were able to examine deeper the specific privacy implications and inferences that can be drawn from the data collected by these devices, due to their expanding popularity. Additionally, fitness tracker users tend to be easier to locate for recruitment compared to smart home users. This accessibility facilitated data collection, participant engagement, and questionnaire-based investigations, ensuring a robust sample size and reliable insights, which was necessary at the time due to resource limitations, which influenced the research decisions. Given the constraints of the study, it was necessary to prioritise and decide accordingly. Focusing on fitness trackers enabled the research team to centralise data collection, analysis, and interpretation, optimising the available resources and maximising the quality of the study outcomes. Research on this topic could provide valuable insights into users' awareness of the privacy risks associated with fitness trackers and assist to the development of privacy frameworks and guidelines for IoT devices.

While the statement regarding questionnaire-based investigations and the availability of fitness tracker users may not be a definitive argument, it serves as an additional practical justification for focusing RQ3 on fitness trackers. The combination of research scope, user accessibility, expertise, and resource limitations collectively supports the decision to concentrate on fitness trackers and provides a solid rationale for why RQ3 was not extended to include smart homes in this particular study.

RQ4 - Can we enhance the awareness of the users regarding the possible inferences that can be obtained from their fitness tracker data?.

The objective of this question is to identify if we can increase the awareness of the users regarding the possible inferences that can be obtained about them from their fitness tracker data, using the results from research question RQ3. We want to examine if the interaction of the users with educational tools that inform them about how their data are collected and shared by their fitness trackers and the possible inferences, increases the level of the users' awareness. The questionnaire used for the exploration of research question RQ3 will be used as a pre-test assessment on the users' knowledge. Then the participants in the experimental group will interact with an educational tool provided to them. After the intervention, the

Alexia Dini

same group of participants will be assessed on their level of awareness of the possible inferences that could be extracted from their fitness tracker data, using the same questionnaire that was used in the pre-test assessment. By comparing the results of the pre-test and post-test assessments, this study will determine whether the interaction with educational tools increases the awareness of fitness tracker users regarding the possible inferences that can be made from their data.

1.3 Research Contributions

Taking into account the research problem and the research questions as defined above,this thesis makes the following contributions:

- We contribute with a list of characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess in order to empower the users to be in control of their personal data and privacy, that has been derived based on an analysis of the state-of-the-art literature, providing an answer to **RQ1**.
- By performing a mapping of user-centric approaches for privacy preservation from the state-of-the-art literature to these characteristics, we also contribute to the research by providing a basis for the design and development of effective user privacy frameworks in IoT, that can be used by researchers for carrying out further research in the area, or by practitioners who can incorporate the characteristics to their platforms or systems, for providing a better protection to their users.
- The present thesis contributes with a conceptual user-centric framework for IoT, “Privacy-EnhAction", that was created based on the characteristics defined from the existing literature and the specific needs for user privacy protection that GDPR requires in the IoT domain. The “Privacy-EnhAction" framework is built on a number of steps, that constitute the processes of the framework and demonstrate how the users can be provided with the functionalities and the tools needed in order to be in control of their personal data created by IoT devices.
- We identify the possible inferences that can be extracted from smart home devices and fitness trackers data by performing a preliminary literature review and by using machine learning techniques in experimental scenarios, providing an answer to **RQ2** and we contribute with a taxonomy of the inferences from smart home devices and fitness trackers.
- We contribute with the development of “PrivacyEnhAction", a privacy tool in the form

Alexia Dini

of a web application, through which the users can analyse data collected from their smart devices or fitness trackers with the objective to be informed about potential privacy vulnerabilities and possible inferences that emerge from the use of these devices.

- We provide the results of a quantitative survey method targeting the users of fitness trackers users aiming to evaluate the level of their awareness regarding the data collected and shared by their devices and the possible inferences that can be made from their data using PrivacyEnhAction as a tool for assessing the user awareness, providing answers to **RQ3** and **RQ4**.
- We contribute with a review of the privacy policies of a number of fitness trackers and smart home devices, in order to look into what data these devices collect, how these data are used, and who can access them, as by helping the users to understand the privacy policies of these devices, they can make informed decisions about which devices to use and what data to share.
- We contribute with *“SpotAware"*, an automated approach that classifies the text of privacy policies from the domains of fitness trackers and smart homes, extracting information for two cases: (a) regarding the eight GDPR user rights addressed in the privacy policy, and (b) about possible data inferences that can be drawn about the user based on the collected data as described in the text of the privacy policy.
- We provide a systematisation of inference groups that include possible inferences or conclusions that could be drawn about the users from privacy policy texts.
- We provide two annotated datasets of 133 privacy policies of smart home devices and fitness trackers for the two cases under study: a) extracting information regarding the eight GDPR user rights present in a privacy policy, and b) extracting information about possible data inferences that can be drawn about the user based on the collected data as described in the text of the privacy policy.

1.4 Thesis Outline

The remaining of this thesis is organised on 10 chapters as follows.

- **Chapter 2 - Background and Related Work:** discusses the background and the related literature review, focusing on the areas of smart home and wearable devices, privacy, the GDPR, and privacy protection in IoT. Following that, the chapter examines inferences from smart home devices and fitness trackers, as in this thesis the areas of focus are these two domains, and investigates user awareness and privacy concerns in

Alexia Dini

those environments, including the analysis of privacy policies as a tool for increasing user privacy awareness. Then, the chapter provides a short introduction to Machine Learning techniques, and concludes with the issues and challenges that research in the area of user privacy protection should address.

- **Chapter 3 - A user-centric privacy framework for personal data protection in IoT:** introduces a number of characteristics that such a framework should possess in order to empower the users to be in control of their personal data and privacy, motivated by a previous work [308]. In this chapter we explain how each characteristic was defined and we proceed with a mapping of user-centric approaches for privacy preservation from the state-of-the-art literature to these characteristics, in order to analyse how they are addressed for user privacy protection and to get an insight regarding the methods and techniques used for the protection of user privacy in various domains of IoT. The details of a generic user-centric IoT privacy protection framework that was created based on the identified characteristics are also presented. Lastly, the chapter summarises all the approaches reviewed providing the findings that were obtained from this part of the research. A significant contribution and novelty of this thesis lies in the suggestion of the conceptual framework based on the identified characteristics that a user-centric GDPR-compliant privacy framework in the context of IoT should possess. Through a Systematic Quantitative Literature Review of the state-of-the-art literature and by integrating relevant concepts, the thesis provides a comprehensive understanding of the essential components and principles that could serve as the basis for privacy frameworks in this domain. This conceptual framework offers a valuable reference point for researchers and practitioners working on privacy and data protection in IoT environments and can assist in the development of user-centric GDPR-compliant privacy frameworks.
- **Chapter 4 - Inference detection in a smart home scenario:** concentrates on two of the characteristics identified in Chapter 3, namely *“CR13:Estimate privacy risks of data collection/inference to users"* and *“CR14:Communicate risks of data collection/inference to users"*, as in this thesis we focus on increasing the user awareness about the possible inferences that can be extracted from their data. We provide details about our methodology that assists in the detection of inferences in a smart home scenario and we describe two approaches we have identified for the analysis of data. We present the results from the Smart Home experiment we set up and the inference types that were extracted, using as a proof of concept the PrivacyEnhAction web application

Alexia Dini

that we have developed to support our research. This doctoral thesis makes a significant contribution to the field by addressing the gap in privacy-aware usage of smart home devices in the GDPR era. The investigation of inferences that can be extracted from data collected from smart home devices adds to the existing body of knowledge by providing concrete examples and insights into the potential privacy risks and implications associated with the usage of these devices. These findings contribute to the understanding of the data inference risks and can inform the development of mitigation strategies and privacy-enhancing measures.

- **Chapter 5 - Inference detection in the fitness trackers domain:** studies further the inferences problem, providing a list of possible inferences that can be extracted about the users from their fitness trackers data, that we have devised from reviewing the available literature. We also discuss the details of the methodology we followed to collect, examine and analyse the data in the fitness trackers scenario under study. Finally the chapter describes how the PrivacyEnhAction tool was extended to include the three fitness trackers to the list of smart devices whose data can be analysed, and we present the results from the fitness trackers experiment along with the inference types that could be extracted, using the PrivacyEnhAction web application to demonstrate the findings. Through empirical investigations, this thesis contributes to the field by successfully identifying the specific inferences that can be extracted from data collected by fitness trackers. By employing machine learning techniques, the potential privacy risks associated with these devices have been exposed, shedding light on the extent to which users' personal information can be inferred from their data. This effort contributes to the understanding of data privacy implications and supports the development of appropriate privacy protection mechanisms.
- **Chapter 6 - Evaluating the impact of PrivacyEnhAction to users awareness:** describes a quantitative study to evaluate the impact that an informative tool, such as the PrivacyEnhAction web application, developed through our research, can have to the awareness of the users regarding the privacy risks and the possible inferences that can be made about them from their data. The exploration of user awareness regarding the inferences that can be made from their fitness tracker data is another significant contribution of this thesis. Through the quantitative study conducted, this thesis has revealed the lack of awareness among fitness tracker users about the potential privacy risks and inferences that can be derived from their data. This contribution highlights an important disparity in user knowledge and highlights the need for education, trans-

Alexia Dini

parency, and user-centric approaches to enhance user awareness and control over their data, while providing novel insights into an evolving and critical area of study.

- **Chapter 7 - Analysis of smart devices privacy policies:** performs a review of how fitness trackers and smart home devices address data collection and sharing and how these are presented in their privacy policies, supplementing our effort to increase the awareness of the users of such devices. This doctoral thesis makes a novel contribution by providing insights on how data collection and sharing practices are addressed and communicated to users. This fills a significant research gap as privacy policies often serve as the primary source of information for users to understand the data practices of these devices.
- **Chapter 8 - Extracting GDPR user rights and inference risks from privacy policy texts:** introduces "*SpotAware*", the approach we propose for the classification of the text of privacy policies from the domains of fitness trackers and smart homes. Our approach contributes to the enhancement of the users awareness as it can extract information regarding how the eight GDPR user rights are addressed in a specific privacy policy, and also information about possible data inferences, such as location or health status, that can be drawn about the user based on the collected data as described in the text. A significant contribution and novelty of this thesis lies in the development and application of the SpotAware approach, which enables the systematic analysis of privacy policies to extract valuable insights, advancing the field of privacy-aware usage of fitness trackers and smart home devices. The SpotAware approach contributes to enhancing user awareness and control over their privacy in the GDPR era.
- **Chapter 9 -The implementation of the PrivacyEnhAction application:** presents the web application that was developed as part of the conducted research, designed to help users analyse the data collected by smart home devices and fitness trackers in order to identify potential privacy vulnerabilities and inferences that can be drawn from their use. The implementation of the PrivacyEnhAction application represents a significant novelty and contribution of this doctoral thesis. This web application has been specifically developed to empower users in analysing the data collected by smart home devices and fitness trackers, with the primary goal of identifying potential privacy vulnerabilities and inferences that can be drawn from their use. The novelty lies in the development of a user-friendly and accessible web application that brings together the insights and findings from the previous chapters of this thesis. The PrivacyEnhAction application serves as a practical tool for users to actively engage in the assessment of

Alexia Dini

their data privacy.

- **Chapter 10 - Conclusions and Future Work:** summarises this thesis by providing the main research contributions. The chapter concludes with directions for future research related to the continuation of this work.

Alexia Dini

Chapter 2

Background and Related Work

Alexia Dini

The research questions introduced in the previous chapter motivate the background and related work review presented in this chapter. First we present how the smart home and wearables domains have developed over time and how these advancements have led to the expansion in the use of smart home devices and fitness trackers. Then this chapter introduces background information related to the notion of privacy and the GDPR, and proceeds with presenting related work in the areas of privacy protection in the Internet of Things, and specifically for the Smart Home and Wearables domains. The chapter continues with presenting how the problems of privacy inference risks in IoT and inference extraction from smart home devices and fitness trackers data have been addressed in state-of-the-art literature. As the main research problem of this thesis has been defined as ***"How can we increase the smart home devices and fitness trackers users awareness over their data and privacy protection?"***, the chapter introduces a review of related works in the areas of user awareness and privacy concerns of fitness trackers and smart home devices users. Then it provides a literature review in the area of privacy policy analysis as a tool for user privacy awareness. In addition, the chapter discusses Machine Learning techniques that are widely being used for privacy preservation.

2.1 Smart Home and Wearable Devices

During the last years, smart homes have become increasingly popular due to the rapid growth and utilisation of IoT. A smart home is defined as "*a residence equipped with a high-tech network, linking sensors and domestic devices, appliances, and features that can be remotely monitored, accessed or controlled, and provide services that respond to the needs of its inhabitants*" [25]. Such devices and appliances include smart thermostats, security cameras, smart washers and dryers, smart vacuums, ovens, microwaves, coffee makers, smart TVs, smart locks, smart lighting, smart speakers, etc. The revolution of devices from simple to smart has been instrumental to the expansion of the smart home, while the advantages that smart home devices bring to the end user can be found in many sectors, ranging from health-related benefits, like Ambient-Assisted Living, to financial and environmental benefits [23, 189].

A key factor in the evolution of the smart home is the ability to detect and recognise the activities taking place in this environment. Machine Learning techniques have been widely used to detect and predict events and activities taking place in a smart home as well as residents behaviour, aiming to provide a better service by analysing the collected user, environmental and context data [56, 176, 220, 262, 342]. Historical records of such events can be exploited to discover patterns in user activities, identify anomalous or suspicious states, or to perform health monitoring and house surveillance.

The progress in smartphone and wearable technologies in the last two decades have made the monitoring and tracking of people's daily activities easier than ever [3], with access to information like walking, running, sleeping, or heart rate being possible by the use of wearable devices. Wearable fitness trackers or activity trackers are IoT connected devices that monitor and track fitness-related metrics like distance walked or run, calorie consumption, sleep quality or heart rate. Fitness trackers consist of sensors such as accelerometers, pedometers, GPS or heart rate sensors, that are used to collect and record the data of the person wearing them, enabling the users to be informed about their fitness and health status [146]. The Global Positioning System (GPS) technology was made available for public use in 1996, and since then it has been used in fitness trackers for the tracking of exercise. The first consumer device with a built-in accelorometer was released by Nokia in 2006, the Nokia 5500 Sport mobile device, as a device to record and track user movement, number of steps, distance, speed and calories consumed. Fitbit was introduced in 2007 as the first wearable device using sensors, starting the new generation of modern activity trackers [268].

Today wearable fitness trackers are low-cost, which in combination with the cutting-edge functionalities that wearable technology brings, has influenced the mass adoption of activity trackers by users [139]. Activity trackers also contain a microprocessor and a communication unit that enables connectivity with a smartphone or a third-party service provider [222]. Wearable fitness trackers come in various types, like smartwatches [336], wristbands [28], waistbands [190], or chest bands [234]. Fitness trackers introduce wearable technologies in users' lives and assist them with personalised fitness advice, by providing actual health-

Alexia Dini

related feedback and by analysing their data to identify patterns. In relation to health benefits, as fitness trackers allow the collection of vital user information, like steps, heart rate, blood pressure, or sleep, this enables the better management of patients assessment and support [257]. Machine Learning techniques have been widely used in fitness trackers for activity recognition [240], the prediction of sleep quality [248], the collection, processing, and analysis of health data [37], personalising fitness plans and the creation of predictions [236] and insights [197] about the users overall health and daily lives.

The extensive increase of smart home devices and fitness trackers use will further add to the amount of user data generated, processed and shared to third parties, allowing the extraction of further insights about the users, bringing to the surface significant user data privacy risks. The protection and control of personal data form important user rights that are recognised by regulations in Europe and worldwide. In the next sections, we discuss these issues in more detail.

2.2 Privacy Definitions

Throughout history, the notion of privacy has been elemental to mankind. Even though privacy developed into an accepted right in the 19th century, it existed long before that [185]. The first written law about privacy was in 1890, known as the “*The Right to Privacy*" [41], where privacy was defined as “*the right to be let alone*" and was expressed as the most inclusive right and the right that is most appreciated by civilised people [6]. Since then, several efforts have been made for the definition of privacy; however, it was not possible to create a common universally accepted definition, as the actual form of privacy varies between societies, economies and cultures.

Some of the most worth mentioning definitions include the one from Westin, who defined privacy as “*the claim of an individual to determine what information about himself or herself should be known to others*" [312]. Westin also defined three levels that affect privacy, the political, the socio-cultural and the personal level [313]. Fried declared that “*privacy is the control we have over information about ourselves* [106], while Gerety states that privacy is “*the control over or the autonomy of the intimacies of personal identity [112]. Another definition comes from Szabo, who defines privacy as “the right of the individual to decide about himself/herself* [273].

Solove argues that no single definition for privacy can be applicable, as there are multiple forms of privacy, related to one another. In order to understand the complex and con-

Alexia Dini

flicting views to privacy, he performed a categorisation of privacy on “*classifications* and “*taxonomies*, dividing privacy into six types which sometimes overlap: (1) the right to be let alone; (2) limited access to the self; (3) secrecy; (4) control over personal information; (5) personhood; and (6) intimacy [267].

According to Parker [221], a definition of privacy should meet three criteria. First, it should fit the data, which means that *data should not be gained or lost through a definition of privacy that is not over broad or too narrow*. A second criterion is simplicity, and this could be achieved by using a list. The example that Parker gives as a definition of privacy at this point fits very well to the IoT data privacy protection problem, as he defines privacy as “*the ability of the individual to lead his life without anyone: (a) interfering with his family and home life; (b) interfering with his physical or mental integrity or his moral and intellectual freedom; (c) attacking his honour and reputation; (d) placing him in false light; (e) disclosing irrelevant embarrassing facts about him; (f) using his name; identity or likeness; (g) spying or prying on, watching or besetting him; (h) interfering with his correspondence; (i) misusing his private communications, written, or oral, or (j) disclosing information given or received by him in circumstances of professional confidence*". The third criterion that a definition of privacy should meet according to Parker is the applicability by lawyers or courts.

2.3 The General Data Protection Regulation

The protection of user personal data in the IoT era has been a challenge for regulators and service providers, as ethical concerns are continuously being raised by researchers [135], driving countries towards the creation of regulations and guidelines, aiming to provide individuals with legal means in order to control their personal data. In Europe, the introduction of GDPR in 2018, aimed to protect the personal data of EU citizens and their rights regarding the use of their data. In this section, we discuss the concepts of GDPR that are relevant to the scope of this thesis.

The GDPR was driven by a rational approach to data protection, that was based on the notion of privacy as a fundamental human right. Under the GDPR, personal data is any information that can identify an individual, and can include a name and surname, a home address, an email address, an identification card number, location data, an IP address, a cookie ID, etc [64]. Furthermore, different types of information that when combined can lead to the identification of an individual also constitute personal data under the GDPR, as

Alexia Dini

well as personal data that have been encrypted, or pseudonyms which can be used to identify a person continue to be considered as personal data [110].

Privacy and data protection have always been a priority for EUs law policy, and as such, GDPR's application comes with the inclusion of seven principles directed to the protection of the user and data privacy and the establishment of user consent for data collection and sharing: (1) Purpose limitation; (2) Fairness, lawfulness, and transparency; (3) Data minimization; (4) Storage limitation; (5) Accuracy; (6) Confidentiality and integrity; (7) Accountability [93]. The principle of purpose limitation relates to the responsibility of the data controllers to use the collected user data only for the purposes strictly defined in their privacy policies, and for no other reason. Fairness, lawfulness, and transparency refer to the obligations of the data controllers to process personal data on legitimate grounds, with the users being made aware of the processing purposes in clear understandable information. The data minimization principle requests that data controllers only use data that are relevant and necessary to the processing purposes and that no excessive data are collected, while the storage limitation principle requires that user data are only stored by data controllers for as long as necessary. In relation to the accuracy principle, data controllers must only store and use data that are accurate, and are required to rectify or delete inaccurate data without delay [307]. Finally, the accountability principle requires from the data controllers to implement suitable mechanisms to protect against data breaches, leaks or unauthorised access.

The GDPR is a crucial regulation for the reinforcement of the users fundamental privacy right, by providing tools to the users in order to control their personal data. This is accomplished with the introduction of eight rights that all EU citizens are entitled to regarding their data. These eight user rights are: (1) The Right to Be Informed; (2) The Right to Access; (3) The Right to Rectification; (4) The Right to Erasure (Right to Be Forgotten); (5) The Right to Restriction of Processing; (6) The Right to Data Portability; (7) The Right to Object to Processing, and (8) The Right to Not Be Subject to Automated Decision Making.

The Right to be Informed requires that before data are collected, the data subject, i.e. the user, *has the right to know how the data will be collected, processed, and stored, and for what purposes* [93]. With the Right to Access, *the data subject has the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, the user must be given access to the personal data and information about: the purposes of the processing, the categories of personal data concerned, the recipients to whom the personal data have been or will be disclosed, the envisaged period for which the personal data will be stored, the existence of the right to request*

Alexia Dini

rectification or erasure of personal data or restriction of processing of personal data or to object to such processing, the right to lodge a complaint with a supervisory authority, any available information as to their source where the personal data are not collected from the data subject, and information about the existence of automated decision-making, including profiling [93].

The Right to Rectification gives to the data subject the right to *obtain from the controller the rectification of inaccurate personal data concerning him or her, also having the right to have incomplete personal data completed* [93]. With the Right to Erasure, *the data subject has the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller has the obligation to erase personal data without undue delay* [93]. The Right to Restriction of Processing gives the data subject *the right to obtain from the controller restriction of processing of his or her personal data* [93]. With the Right to Data Portability the data subject *has the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and has the right to transmit those data to another controller* [93].

The Right to Object to Processing gives the data subject *the right to object at any time to processing of personal data concerning him or her* [93], and finally, with the Right to Not Be Subject to Automated Decision Making the data subject has *the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her* [93].

The GDPR is highly concerned with fitness trackers and smart home devices, as their functionality involves the use of personal data, and as such they have to comply with its directions. Since transparency is key in the application of GDPR, it is essential that fitness trackers and smart home devices users become aware about how their personal data are processed [29], as fitness trackers collect enormous amounts of highly sensitive personalised body, health and fitness data, like activity, steps count, temperature, sleep patterns or calories burnt, using embedded sensors such as pedometers, accelerometers, GPS, heart rate monitors, altimeters, etc. [323], while smart home device ranging from a smart bulb to a smart thermostat collect personal data continuously in the smart home.

Alexia Dini

2.4 Privacy protection in IoT

2.4.1 Privacy protection frameworks

The heterogeneous structure of IoT devices adds more weight to the need for efficient privacy protection in the data management process making the requirement for efficient privacy protection frameworks even more notable.

IoT devices process and analyse the collected data in order to provide smart services to the users, usually combining these data with data from other applications with the aim to provide a better service. However, the handling of personal user data which can lead to inferences when combined with other data, such as profiling, is prohibited in GDPR (Article 9). Previous works have focused on the prevention of data inferences through the use of inference prevention techniques like probabilistic models or learning algorithms [288], by enabling the users to define what cannot be inferred from their data [46], or by performing privacy risk analysis enabling the users to understand the privacy risks of each privacy choice using harm trees [77]. Other techniques involve data transformation, which has been the focus of many analyses since the GDPR was introduced. Under the GDPR directions, it is acknowledged that data transformation techniques are essential for the protection of user data, as they provide a way to boost the protection of privacy. Examples from the literature involve the application of anonymisation to data before it is released [231], the use of differential privacy in a smart metering use case [26], or data masking, such as perturbation, randomisation, or quantization [294].

In [288], a framework is presented that uses a personal data manager that can assist in the management of third party data request in the context of IoT, using an inference risk calculation process associated to data disclosure. In [130], the authors present UPECSI, a solution that allow users to enforce their privacy requirements before any personal data are uploaded to the cloud. Through the proposed tools, the end-users can protect their data, while cloud service providers can easily integrate privacy processes by providing users with a transparent user interface that enables them to specify their privacy settings accordingly. PISCES, a Privacy by Design (PbD) framework, is proposed in [104], that provides private data management for IoT through a rigorous disconnection between data providers and data controllers, allowing the user to know by whom and for which purpose her data are being used. A policy-based and privacy-enabling framework for informed consent in IoT is proposed in [207], where usage control policies are utilised for the regulation of access to personal

Alexia Dini

Table 2.1: Summary of related work for privacy protection in IoT

Area	Cit.	Proposed
Privacy protection Frameworks	[288]	Probabilistic models, learning algorithms
	[46]	Users can define what cannot be inferred from their data
	[77]	Privacy risk analysis using harm trees
	[231]	Data is anonymised before released
	[26]	Differential privacy
	[294]	Data masking
	[288]	Enforcement of user privacy preferences
	[130]	Privacy enforcement points
	[104]	Private data management
	[207]	Usage control policies
	[244]	Smart object check compliance of user privacy preferences
	[77]	User interface displays impact of user privacy choices
Smart Home Privacy protection	[180]	Development of privacy protection standards
	[340]	Blockchain based privacy protection scheme
	[325]	Data encryption using DES encryption and LSB algorithm
	[126]	Sample data analysis and supervised learning techniques for timestamp series data
	[232]	k^m data anonymization technique
	[324]	Encryption and information hiding methods used
Wearables Privacy Protection	[199]	User interface for user privacy preferences
	[36]	Provides best practices for manufacturers
	[152]	Local differential privacy
	[14]	Data anonymization technique
	[150]	Analysis of third parties communicating with devices

Alexia Dini

data, that can be customised according to the specific needs of the user and the context. In particular, the users can define their security policy rules, which are stored in a user-centric Security Gateway. The users can define specific policies for each Service Provider managing access to data. A privacy framework for IoT where smart objects users have the ability to specify their privacy preferences is presented in [244], where the smart objects check the compliance of the privacy preferences of the users, assisting the aim of the framework which is to enhance the privacy of IoT users. In [77], a framework enhancing individual control over personal data is introduced, where the user can indicate her privacy choices, view the impact of those choices on the available user interface and take the appropriate decision to revise them or not, based on the privacy risks present.

2.4.2 Privacy protection in the Smart Home

In the smart home a number of devices are connected aiming to provide an effective, comfortable and energysaving environment. Along with the benefits that the smart home has brought to people's lives, it has also caused a huge expansion in the creation of data, leading to the creation of a number of privacy risks to the user and the data. Smart home privacy protection has been the focus of the research in the last decade, aiming to minimise the privacy risks and make the smart home a safer environment.

The development of privacy protection standards for the smart home is presented in [180] aiming to provide support for data security governance in this area and to increase the level of social privacy protection. In this work, the standards are divided into three phases, creation, exploration, and expansion, where critical technology, auxiliary management, test and certification, and device application are incorporated. The privacy problem of the bi-directional interaction between smart homes and the power control centre is the focus of the work in [340]. As smart meters collect fine-grained real-time electricity or water consumption data, the procedure to upload these data to a control centre can lead to the leakage of these data, as privacy problems may occur like single point failure, malicious tampering of data, etc. In this work, a privacy protection scheme for smart meters based on the blockchain technology is proposed, that ensures that the users privacy is protected, providing confidentiality, with lower computing and communication costs. The problem of malicious adversaries intercepting transmitted smart home data is targeted in [325] through a smart home privacy protection method that combines DES encryption and the improved Least Significant Bit (LSB) information hiding algorithm. In the proposed method, the smart home data is encrypted using DES encryption and then the LSB algorithm is used to hide the cipher-text, providing a double protection scheme to the data.

In the smart home, the daily activities of the residents can be determined through the analysis of the timestamp series of the data. A method to hide the patterns of the daily routines of smart home residents is presented in [126], where by using sample data analysis and supervised learning techniques this problem is overcome providing effective privacy protection, low energy consumption, low latency and strong adaptability. In another work, an architecture is presented for a fog enhanced smart home environment that preserves the privacy of the users when their data is shared with third parties [232]. Here, the authors propose a k^m data anonymization technique for the prevention of data breaches. A privacy protection scheme for the smart home is proposed in [324] that is based on information hiding. Using

Alexia Dini

Machine Learning techniques, the smart home data are classified into sensitive data and non-sensitive data, a process which can be controlled according to the users' preferences. Then, the sensitive data are transmitted after they are processed by a combination of a method that uses both encryption and information hiding. As the proposed scheme combines both the ability of the users to express their privacy preferences and the encryption with information hiding, the privacy of the data is greatly enhanced.

2.4.3 Privacy protection in Wearables

Fitness wearables include devices like sport watches, smart watches, wristbands, chest straps and other smart gear, that monitor and track the number of steps we take every day, how many stairs we climb, the number of hours we sleep every night, or the quality of our sleep, among others. Studies have shown that smartphone users are most likely to own a fitness wearable [24], while compatible Fitbit devices enable the users to make contactless payments, providing additional services. Data collected by wearables can be exploited in the pursue of inferring information regarding bodily activities like walking or running [58], while smartwatch data have been successfully used for the recognition of user eating [283] and drinking activities [219], or smoking [275].

Since the essence of wearables and fitness trackers does not usually allow a high level of interaction between the device and the users, a user interface is proposed in [199] for capturing the privacy preferences of the users in each application they use. The presented GUI aims to educate the user about data access requests and protect her personal data. The privacy vulnerabilities and threats of using fitness trackers, and in particular the Fitbit smartwatch, are explored in another work [36], by analysing the device features and potential privacy risks. The authors present a list of actions to diminish these vulnerabilities and they propose a number of best practices for wearables manufacturers to provide balance between functionality and privacy protection.

As the sensitive information collected by fitness trackers needs to be protected, a method for accumulating and processing health data in a privacy-preserving way is presented in [152]. Local differential privacy is being used adopting a sampling-based data collection scheme that accomplishes an important advancement in accuracy than simpler solutions, providing better privacy protection on the data collected. An anonymization approach is proposed in [14] to protect the privacy of the users data from smart health devices, by generalising pivotal, data aiming to make it arduous to re-identify a user. According to the authors, the

Alexia Dini

Table 2.2: Summary of related work for privacy inference risks in IoT

Area	Cit.	Data used	Inference Risks
SH	[342]	Smart home data combined with user social network profile	User location
	[90]	Patterns of water consumption	Occupancy, vacation, periods, activities
	[213]	Motion sensor data	Single or multi-user presence in a house
	[326]	Motion sensor data	Occupancy, user identity
	[194]	Electricity data	User routines
	[342]	Linking data from different smart devices	Personal owner information
	[162]	Smart devices data	User habits, preferences
	[164]	Smart speakers data	User habits, preferences
	[166]	Accelerometer sensors data	User habits, preferences
	[33]	Smart meter energy data	User routines and habits
	[50]	Location information, sensor data	Various
FT	[164]	Accelerometer sensors data	Activity, behaviour, location tracking
	[323]	Pedometer sensors data	User routes
	[196]	Elevation data	Location, frequently used places
	[240]	Fitness tracker data	User activities

results of this technique demonstrate that with a small compromise on computational cost and data retention, the solution is effective for privacy protection. An analysis of the third parties that communicate with fitness trackers and their associated smartphone applications is presented in [150], where any unexpected - from the privacy point of view- third parties are identified. The aim of this work is to urge the users to study the privacy policies of devices before purchasing them to learn more about what personal data are being shared.

Table 2.1 provides a summary of the related work discussed in Section 2.4. While the work in [230] is occupied with user privacy awareness in the area of wearables and IoT services by presenting a framework that could be used as guidance to developers and service providers in order to integrate privacy risk user awareness in their products, no other work to the best of our knowledge has been involved with raising user awareness in relation to the inferences that can be extracted about the users from their fitness trackers data.

Alexia Dini

2.5 Privacy Inference Risks in IoT

Artificial intelligence (AI) is progressively being used in numerous fields to draw inferences, insights and predictions about the behaviour, preferences and lifestyle of people [309], through its ability to process and evaluate enormous datasets. These inferences are unverifiable and open the door for discrimination and biased decision-making. In the IoT domain, the risk of undesired inferences drawn from personal data becomes higher as the number of connected devices used increases. Most users are not aware of the extent of data collected by those devices, making it confounding to understand that these data can reveal more information about them. In this section, we present the related work in the area of inferences as a privacy threat in IoT, and in particular, inferences that can be drawn from smart home devices and fitness trackers. Table 2.2 summarises the related work for privacy inference risks in IoT presented in this section.

2.5.1 Data collection in IoT

In the context of the IoT, data collection is a central feature that enables the functionality of smart devices. These devices are equipped with sensors and communication capabilities that continuously collect different types of data. The data collected in IoT environments can be diverse and extensive, including sensor data, coming from a wide range of sensors, such as temperature sensors, humidity sensors, motion sensors, light sensors, and more. These sensors collect environmental data, like temperature levels, ambient light conditions, air quality, and presence detection. Additionally, IoT devices collect data related to user interactions, including button presses, voice commands, touch gestures, and other input methods. These data can provide insights about the users' preferences, behaviour patterns, and usage habits. Smart devices often track and record energy consumption data, allowing users to monitor and optimise their energy usage, including data about electricity consumption, water usage, and other resource utilisation information. IoT devices that are equipped with cameras or microphones can capture multimedia data, like images, videos, or audio recordings. These types of data can be used for various purposes, such as surveillance, remote monitoring, or voice-controlled interactions. Furthermore, smart devices can collect contextual data related to the user's location, time, and surrounding conditions, that include GPS coordinates, timestamps, weather conditions, or proximity to other devices or objects. Fitness trackers continuously monitor and track users' physical activities, such as step count, distance travelled, calories

Alexia Dini

burned, and sleep patterns. These data provide insights into users' daily routines, exercise habits, and sleep patterns. What's more, some fitness trackers are equipped with sensors that monitor heart rate, blood pressure, and other biometric measurements, which can reveal valuable information about the users' well-being, stress levels, and overall health.

2.5.2 Information inference as a privacy threat in IoT

In the literature it has been shown that seemingly harmless data from smart devices can be used to infer eminently personal information about the users [162]. Machine Learning techniques and big data analytics have been used for drawing vigorous inferences from apparently harmless data or identified behaviour, compromising a basic privacy law, which is to allow a person to control who knows what about them [134]. Similar techniques are also used for making predictions about people's private lives, behaviours, habits and preferences, establishing the perfect conditions for discrimination, prejudicial and intrusive decision-making against the people involved [309], creating a crucial threat to user privacy. Recently, these privacy related concerns have expanded from personal worries to social issues, as "anonymised" fitness tracking data from Strava, a widely used application for tracking activity and exercise, were released in the form of an "anonymised" heat map. The company mapped its accumulated activity data of two years in order to display the most visited areas in the map. However, US secret war zone locations and military bases were highlighted as soldiers habitually upload their fitness tracking data to Strava, creating a massive security threat as sensitive government and military sites were exposed [314].

In the domain of IoT, inferences are personal information that are not consciously provided by the users themselves, but extracted by data controllers or other third parties from given data. This is a common approach in the area of Machine Learning; still inferences can be obtained without the use of advanced techniques. A "current" example of an inference that can be extracted without the use of Machine Learning or other advanced techniques, relevant to the Covid-19 pandemic, is the following: A person could be thought as having the virus, if that person has travelled to a heavily infected area during the recent weeks. The inference being made here is not a proof that a person has been tested positive for Covid-19, but an indication of the possibility of infection [264].

The problem of undesired inferences is more evident in IoT due to the increasing amount of data generated and the available data analysis techniques, and they constitute a major threat to the privacy of the users. The subject of privacy protection has been a challenge for

Alexia Dini

the researchers since the beginning of the digital age [104]. Today, the EU data protection authorities acknowledge the need for the assurance of personal data protection, and in particular the processing of health related data, which is generally prohibited under GDPR Article 9[1]. As inferences are only predictive and indicative, they may be inaccurate and unverifiable. Nevertheless, they contribute to the creation of user profiles by companies and third parties and could potentially jeopardise people's basic rights and privacy, as the more data that are collected and associated with a user, the more inferences can be made about that user.

2.5.3 Inferences from Smart Home devices

In the era of smart homes, sensitive data are recorded and transmitted to multiple service providers. In most cases, such data are used to provide high-quality, useful services to the citizens. There are situations however, where the same data can reveal sensitive information about the user, if obtained by an unauthorised party. Unintended inferences have become the biggest threat to privacy, mainly as a result of the development of sophisticated Machine Learning techniques [309].

Devices ranging from a smart bulb to a smart thermostat collect data continuously in the smart home. The risk of undesired inferences drawn from personal data becomes higher as the number of smart devices used increases. For example, smart metering data can disclose when a person is at home, the frequency a user watches TV, cooks, sleeps or goes on holidays. This type of information is invaluable to third parties like employers, insurance companies, or marketing companies, who are very interested in gaining access to it [224]. The EU GDPR has arrived to give control to the users over their data and the protection of their privacy, therefore the need for purposeful tools that allow the users to be informed and understand the privacy risks of using a smart home device is imperative [160].

A key factor in the evolution of the smart home is the ability to detect and recognise the activities taking place in this environment. Machine Learning techniques have been widely used to detect and predict events and activities taking place in a smart home as well as residents behaviour, aiming to provide a better service by analysing the collected user, environmental and context data [342]. Historical records of such events can be exploited to discover patterns in user activities, identify anomalous or suspicious states, or to perform health monitoring and house surveillance. Furthermore, when smart home data is combined with data from other sources like social networks, further privacy vulnerabilities become

[1] https://gdpr-info.eu/art-9-gdpr/

Alexia Dini

apparent, especially when context information is considered. For instance, the social network profile of a user can be used to predict the user's whereabouts at a specific time conforming to knowledge released by friends [342]. This is possible, due to the linkable published content of users which may include sensitive information for each other, allowing for the inference of users sensitive locations based on the social relationships of a user.

The patterns of water consumption from a smart water meter can be used to infer house occupancy, vacation periods, or even which rooms are being used and when, and also occupants activities like using the toilet, bathing, etc. [90]. Data collected from motion sensors may contain contextual information like date, time of day or location, that can be exploited in order to infer single or multi-user presence in a house [213]. Machine Learning models can be applied to motion sensor data to infer which particular rooms are occupied in a house or even the occupants identities [326]. A framework that analyses the electricity consumption of a house to unveil household characteristics relating to the economical or social status of the family, the number of occupants, or the home appliances being used is presented in [95]. In [194], electricity data from smart meters is used to analyse the trends between different customers, obtaining information such as personal details about families, like the time they wake up and have breakfast, when they go to work, if someone stays at home during the day, etc. Most of these works do not aim to utilise the detection of the inferences towards the user benefit, but for the improvement of various services.

The data privacy vulnerabilities that are created from the linking of data from different smart devices are presented in [342] by validating how this linking can reveal personal information about the owners. An analysis on how sensitive information can be captured from data generated by smart devices and be used to make additional inferences about habits, preferences, and so on is described in [162]. The same author presents his findings in relation to the extraction of inferences from human speech and other sounds collected by smart speakers or other smart devices that record audio [164] and accelerometers [166]. The findings regarding the likelihood of the improper use of smart meter energy consumption data in order to expose the routines and habits of people are presented in [33]. The possibility that location information and sensor temporal sampling can be used to compromise a smart home user's privacy is explored in [50]. A privacy model is proposed and preliminary analysis shows that it is possible to compromise the privacy of a user through inferences drawn from their data.

Occupancy monitoring is an area that has attracted a lot of research interest, aiming to provide insights for improving energy or water consumption in buildings, cut down expenditure and enhance performance. However, in the smart home scenario, if the results of

Alexia Dini

occupancy monitoring are misused then the privacy of the occupants can be compromised. A framework that can prevent occupancy detection in a smart home through the use of adversarial Machine Learning techniques is presented in [263]. The framework aims to enhance the protection of the users' privacy, offering customised user privacy preferences. In [261], heterogenerous sensors and Machine Learning techniques, such as Random Forest and Linear Discriminant Analysis (LDA) are exploited, in order to estimate the number of persons in a room, while in [326] it is shown that by using smart meter and motion sensor data, occupancy related inferences are possible, like the number of occupants or even their identities.

Activity detection in smart homes is another area that has received a lot of attention, mostly for Ambient Assisted Living (AAL) applications, where the user data are being used to provide a better service. Yet, most of the works do not acknowledge the privacy risks this process entails for the user. A methodology to infer the activity profiles of households using smart meter data is proposed in [316], while in [332], the authors exploit smart meter data in order to discover activity patterns using Machine Learning techniques for healthcare applications, aiming to provide assistance to the elderly living alone when anomalous activities are detected.

2.5.4 Inferences from fitness trackers data

As the use of fitness trackers is increasingly growing among users, the amounts of personal data created are enormous. These data are often handled by third parties or service providers for the provision of the relevant functionalities. However, in the literature it is reported that these user data can be exploited by third parties for the extraction of personal and sensitive information about the users through various techniques, such as Machine Learning algorithms, data mining techniques, predictive filtering, etc. [287].

The use of accelerometer sensors embedded in wearable devices is exploited in [164] presenting a number of inferences that are possible from analysing the data collected by such sensors. The identified inferences include activity, behavior or location tracking. The authors suggest that their findings should be used as a caution to customers and a cause for action to developers and organisations. The possibility of inferences from pedometer sensors that are used to count steps is studied in [323]. The possibility of inferring the user typical routes, for example going to a coffee shop or a grocery shop, is computed by utilising the steps per minute data from the user's fitness tracker. The Euclidian distance between the steps-tracked sequence and the path query sequence is used to set a threshold value, and

Alexia Dini

as long as this fluctuates, then the user route can be inferred with an accuracy of almost 50%. The elevation data from fitness trackers are used by the authors in [196] to predict the location path of the users, using natural language processing computer vision for the representation of data, and Machine Learning and deep learning-based techniques to predict and infer personal information, such as frequently visited places. A case study based on fitness trackers is presented in [287], where a model for inference prevention is built using a Bayesian Network, that computes the risk of inference attacks from the combination of known data about users.

A study on the privacy vulnerabilities of fitness trackers is presented in [240], where Machine Learning techniques are exploited for the analysis of data from these devices in order to make meaningful inferences about user activities. The results show that it is possible to track users and their activities from their fitness tracker data, creating a threat to their privacy. The possibility of privacy leakages from Bluetooth Low Energy (BLE) communication between fitness trackers and smartphones is examined in [76]. As the BLE traffic of fitness trackers seems to be correlated with the intensity of the user activity, the authors show that it becomes possible for a malicious listener to infer the user's activity, by analysing the BLE traffic analysis. They also present their findings regarding the possibility to identify a user by analysing the BLE traffic of her devices, which can depict the unique way a person moves.

The overlooked security and privacy challenges in wearables is the focus of the work in [34], where the authors identify a number of inferences that can be extracted from sensors data. According to the authors, fitness trackers become an appealing source of interest for cyber criminals, whose attacks may gain access to users bio-metric data, enabling identity theft, location information which is a major privacy threat, or accelerometer data that can be used to infer user activities. Subsequently, the authors recommend that further research is needed for the consideration of privacy requirements early in the design of fitness trackers and wearables in general.

Fitness trackers record the number of steps taken every day by the users, as a measure of their activity level. Activity can be classified using the step index in Table 2.3 that has been proposed by Tudor and Basset to describe the physical activity in adults based on pedometer readings [292]. No or low physical activity is the root behind ill health [305], therefore knowledge of this kind of information could be an indication of possible health problems. Information like daily walking step count may potentially reflect peoples stable lifestyle and habits or whether someone is at a lower or higher risk of all-cause mortality [245]. Low levels of daily activity could indicate that the user may be suffering from health problems.

Alexia Dini

Table 2.3: Activity levels and steps indices [292]

Activity Level	**No of steps**
Sedentary	Less than 5,000 steps per day
Low active	5,000 to 7,499 steps per day
Somewhat active	7,500 to 9,999 steps per day
Active	More than 10,000 steps per day
Highly active	More than 12,500 steps per day

Alexia Dini

This information can be used by an interested third party, such as an insurance company, to increase health insurance premiums based on the identified behaviour, for example when the user does not lead an active or healthy lifestyle.

Activity data can also be used to infer religion. This can be applied particularly for the case of the Orthodox Judaism religion, as on Saturdays believers engage in restful activities to honour the day according to their religion. Even though for most people Saturday is an off-duty and leisure day, if it is observed from the fitness tracker data that the user is usually very active on most days but not on Saturdays, then this could be seen as an indication - not a proof - that the person may be Jewish [67]. Religion could also be inferred by the time the person wakes up in the morning, since Muslims wake up earlier during Ramadan [302]. Religious or philosophical beliefs are considered as sensitive personal data and could be used in a discriminatory way against a user if obtained by a third party; for example a potential employer.

A number of fitness tracker devices collect the user's VO2Max (cardio fitness level) values. This measurement is thought to be the best indicator of cardiovascular fitness. Monitoring VO2Max over time can assist in establishing whether a person is getting fitter or losing their fitness. Research in the area has shown that low cardio fitness levels are linked with cardiovascular disease, while higher levels are correlated with many health advantages [98], [133]. Therefore, a declining or increasing VO2Max can be used as an indicator of the overall fitness of the user.

Heart rate data collected by fitness tracker devices are very important and include a treasure of information about our bodies. According to European data protection bodies, heart rate information constitutes part of health data, while under the GDPR, *"personal data concerning health should include all data pertaining to the health status of a data subject which* *reveal information relating to the past, current or future physical or mental health status of the data subject"*[2]. As such, health data including heart rate measurements are considered as a special category of personal data.

Insights about heart rate measurements can assist in observing and understanding one's fitness level, but also to identify possible health problems. The values of resting heart rate, i.e. when the person is sitting and is calm, relaxed and not sick, varies between 60 beats per minute to 100 beats per minute for adults[3], therefore a resting heart rate of more than 100 beats per minute is considered high, while a heart rate of less that 60 beats per minute is considered low. A resting heart rate that is below the normal range could be due to a number of reasons. It is a normal situation for a person that is an athlete or a fit and young adult, or it can happen as a side effect of taking a specific medication or from a health condition, such as bradycardia [141], [188]. To that end, having a low resting heart rate could indicate that the user is an athlete, she may suffer from bradycardia or is under medication at the time of the readings. An elevated heart rate could be due to a health condition, exercising at the time of the readings or heavy alcohol consumption, consequently the inferences that could be extracted about the user from these data are that the user may be suffering from a heart condition, or could be an alcoholic [68], [10]. The users could face discrimination or increased premium rates, if third parties got hold of such data.

Location data can reveal individual mobility patterns; when combined with fitness activity information, it may reveal the areas a person mostly works out or even that person's home or work address [217]. Furthermore, users' fitness activity could reveal their behavioural patterns, including the hours when they are usually away from home. The privacy risk is that if this information falls in the hands of a malevolent third-party, then the personal or home safety of the user could be jeopardized. The GDPR acknowledges the location datas unique position as identifiable information by making it part of its definition of personal data in Article 4[4]. In the absence of location privacy protection, aggressors can exploit this gap to carry out a variety of attacks. These attacks may include: (i) undesired advertising to users of products near to the user proximity, (ii) physical attacks and harassment or user profiling and tracking, when location data can be used to infer other sensitive information, such as state of health, personal habits or professional duties, (iii) political, religious, sexual persecution and

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679from=GA

[3] https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse

[4] https://gdpr-info.eu/art-4-gdpr/

Alexia Dini

discrimination, in which a person's location is used to restrict his or her freedom [71], (iv) planned break-in according to the times the user is away from home, (v) stalking.

Sleep tracking is a feature that is supported by most fitness tracker brands, where by using heart rate sensors and accelerometers for movement monitoring, sleep can be detected automatically. Science has long recognized the importance of sleep to physical well-being. People who get less than six hours of sleep have a threefold increased risk of high blood pressure versus those who sleep more, and women who get less than four hours of sleep have a twofold increased chance of dying from heart disease than those who sleep longer [204]. Moreover, in research it has been reported that lack of quality sleep is associated with diabetes, obesity, and cancer, as well as worse memory and mental health. On the other hand, getting too much sleep is also associated with health problems. Since sleep is fundamental in people's prosperity and physical and mental wellness, lack of sleep and bad quality of sleep have been proven to be linked with health problems, reduced cognitive functioning, bad mood, and reduced productivity [52].

Furthermore, the extraction of users' sleep patterns from data collected by fitness trackers can be used for user profiling. These user profiles can potentially be exploited by marketing or pharmaceutical companies for targeted advertising, when combined and correlated with other data, like heart rate or interests [40]. A user's personal safety could also be at risk since by tracking sleep patterns, information about when the user ordinarily has the deepest and lightest sleep becomes available, as some fitness trackers collect information about sleep stages. Inferred wake up times may be used by third parties, such as marketing companies, and the user could be targeted for unwanted advertising, since people have better working memory accessibility in the morning close to the time they wake up [299]. Additionally, the average percentage of light sleep, deep sleep and REM sleep stages that can be inferred can reveal further insights about user focus capability, mood, memory, use of possible medications like antidepressants, anxiety, depression, etc., while it can be concluded that people who are sleep deprived are also more likely to make errors and omissions, and could then possibly be discriminated against by current or potential employers.

2.6 User awareness and concerns on IoT privacy

Various studies in the existing literature engage in collecting and analysing the opinions and perceptions of the users of smart home devices and wearable devices regarding the protection of their privacy and the possible risks from the exposure of their personal information without

Alexia Dini

their awareness or consent. In Table 2.4, the summary of the related work for user awareness and concerns on IoT privacy presented in this section can be found.

2.6.1 User awareness and privacy concerns of smart home devices users

The security and privacy concerns of smart home devices users is the focus of the work in [335], where the authors use semi-structured interviews with residents of smart homes in order to understand their privacy views and concerns, as well as the users' awareness over smart home security issues. Through the study, the authors came to the conclusion that the participants were more worried about physical security issues, like acts that can compromise their safety at home, rather than privacy issues, and in some cases the participants were unconcerned about the privacy issues of their smart homes. Another important finding of this work is that even though a number of participants showed awareness of some privacy issues of the smart home, they expressed limited concern.

Another study employed a questionnaire instrument [259] to examine the impact of users' personal factors, such as awareness and trust, on smart home acceptance. The results of this study showed that the perceived smart home privacy risks can prevent the user from trusting smart home devices, which in turn influence their intention in using them. Additionally, the study confirms that users' awareness on the associated privacy risks negatively impacts the users' attitude towards using smart home devices.

In another work, the authors investigate the users' perceptions of smart home devices privacy risks by using semi-structured interviews [341]. This study reports that users are apprehensive of the privacy risks of devices like thermostats or smart lights, while they are not aware of the risk of inferences from their data, such as sleep patterns or home occupancy. Furthermore, according to this study, the participants trust that leading brands in the sector provide adequate privacy protection. The user privacy and security concerns of smart homes are the focus of the study in [344], where results of semi-structured interviews show that the users are worried about their data collected by smart home devices. The participants are conscious about sensitive data like bank details, times of absence, personal preferences, hacker attacks and data abuse.

To better understand how users comprehend the security and privacy risks of the smart home devices they own, a survey was carried out [209] on users and non-users of smart home devices. In the study it is reported that those participants that are aware of the vulnerabilities of IoT devices, also consider that the privacy protection of smart home devices is essential.

Alexia Dini

Table 2.4: *Summary of related work for user awareness and concerns on IoT privacy*

Area	Cit.	Method	User privacy concerns
SH	[335]	Interviews	Physical security issues
	[259]	Questionnaire	Awareness and trust
	[341]	Interviews	Users apprehensive of the privacy risks of some devices
	[344]	Interviews	Bank details, times of absence, personal preferences, hacker attacks, data abuse
	[209]	Survey	Some users apprehensive of the privacy risks
	[60]	Online reviews analysis	Tracking, data hacking, 3rd parties access personal data, devices always listen to conversations
FT	[172]	Survey	Disclosure of financial information, location, stalking, physical harm
	[173]	Interviews	Disclosure of medical information
	[62]	Survey	3rd parties access personal data, devices collecting too much information, activity monitoring
	[100]	Survey	3rd parties access personal data, data used against the users
	[343]	Survey and interviews	Users do not perceive data collected from fitness trackers as sensitive
	[138]	Survey	3rd parties access personal data, data used against the users

On the other hand, those participants that believe that the privacy protection of smart home devices is not important are also those people who do not buy IoT devices because of the cost, and not due to privacy issues. The work in [60] also concentrates on the privacy risks of smart home devices. Here, the authors performed an analysis of online reviews of consumers of smart home hubs in their effort to extract the privacy concerns from a user-centric perspective. While one third of the reviews were general, from the rest of the available reviews a number of user concerns could be identified, such as the worry that these devices always listen to the conversations, tracking of users, their actions and preferences, storage of conversations, lack of security, the potential of private conversations to be hacked, or the possibility that their information will be disclosed publicly.

Alexia Dini

2.6.2 User awareness and privacy concerns of fitness trackers users

User concerns related to personal data privacy risks are investigated in [172], where by using a survey with a number of data exposure scenarios in their study, the authors assess user concerns and their results indicate that privacy is at the top of the users' worries when using wearables. On the other hand, the authors have also observed that the users are eager to accept any privacy related risks, if they consider that the benefit associated with that risk is significant to them. Furthermore, the users' main concerns identified in this study include (a) the disclosure of financial information, which is a user concern related with any possible costs that the user may suffer from the disclosure of stored financial information on their fitness trackers, and (b) location tracking, stalking and physical harm as the result of the use of GPS technology on some wearables. The results of this work provide insights related to how the users of wearable devices discern personal data disclosure.

The user understanding of the privacy and sensitivity of the data collected by wearable devices is studied in [173]. Using a qualitative research approach to collect data through themed interviews, the study found that overall the participants do not consider the data collected by activity trackers to be private, except in the cases when such data are combined with identifiable information, like name and address. On the other hand, the participants considered health information stored in medical records very sensitive and private. As such, the disclosure of medical information has been identified as a user concern, since users are worried that third parties like banks, insurance companies or employers could potentially benefit from such data when taking decisions regarding loans, insurance rates, hiring new staff, promotions, etc.

The factors taken into account in the privacy calculus of wearable fitness devices are analysed in [62], where a research model is developed based on the privacy calculus theory and uses a survey administered to fitness trackers users in order to examine if there is a relationship between the users' intention to disclose personal data and to continue using the wearable device. The results of the survey led to the observation that the users are more likely to continue using the device if the perceived benefits are higher that their privacy concerns. Identified privacy concerns include the possibility that third parties could gain access to users' personal data, the likelihood that the devices collect too much information about the owners and activity monitoring.

The users' understanding of the data collection in fitness trackers and their privacy concerns are studied in [100]. The authors have used an online survey where current, former or

Alexia Dini

non-users of fitness tracking applications from the EU and USA have participated in order to determine how the different groups comprehend the sensitivity of the data that are collected by these devices and what specific concerns they have in relation to their privacy. The main finding of this study is that users who generally feel insecure about their data privacy online are also more likely to be worried and concerned about the protection of the privacy of their data collected from fitness trackers. User privacy concerns identified through the survey include the likelihood that third parties could gain access to their personal data and that their data could be used against them. In the work in [343], the authors employ a survey and semi-structured interviews with current users of fitness trackers in their effort to gain an understanding on the advantages and disadvantages that users perceive from their interaction with these devices. In general, the participants indicate that they have low levels of concerns regarding their privacy and that they consider that the benefits of using a fitness tracker exceed any disadvantages. The outcome of this study according to the researchers is that the users do not perceive data collected from fitness trackers as sensitive, they are not aware of possible threats and they are inclined to share their personal data, like heart rate or step count, as they feel that the privacy risks are low. A survey with the goal to investigate the likeness and dissimilarities of fitness trackers users' privacy attitudes from USA and Germany showed that the weight of a number of user privacy concerns varied considerably between the two groups [138]. The introduction of the GDPR in the EU was the driving force for this study, and it has been shown that the European users are using their GDPR rights and have become more responsible of their data. Examples of the identified user privacy concerns include among others the possibility that third parties could gain access to the users' personal data, or that their data could be used against them.

An analysis of how fitness tracker users understand the privacy inference risks affiliated with the use of these devices is presented in [302]. Through the use of a longitudinal study, an online survey and interviews with the participants, the authors come to the conclusion that the participants are apprehensive of the types of information that might be inferred about them from their fitness trackers data. The authors go one step further and suggest that one solution to protect the user's privacy is to offer better data minimization procedures by dropping centralised data collection and by decreasing the granularity of the data collected and sent to the data provider.

Alexia Dini

2.6.3 Information inference threats in other domains

Information inference is possible in different domains and with different types of data. Researchers have pointed out that it is possible to exploit someone's personal information like birth date and place of birth to infer their social security number [2]. In the gaming domain, personal information such as age, gender, emotions, interests, habits and personality traits can be inferred through the analysis of in-game behaviour and collected gaming data [167]. Voice recordings have been widely used in research for the extraction of information about the user, such as geographical origin, gender, age, health status, mood, emotions, personality traits, etc. [167]. Names and contact information have been successfully linked to public profiles containing medical information such as procedures and diseases, as well as information like gender, date of birth and postcode, making the identification of profiles possible [272]. Personality characteristics and friendships networks can be reliably predicted using call logs [78] and empirical data like Bluetooth proximity, app usage or phone status [88], or location [85]. Facebook behaviour records, such as “likes" have been used for the prediction of a number of sensitive personal features, such as political and religious beliefs, sexual preferences, ethnicity, alcohol or drugs use, age, and more [156]. Research shows that eye tracking data can provide rich and sensitive information about a person and such data have been used to extract information with respect to that person's biometric identity, gender, age, ethnic origin, personality characteristics, drug use, emotional state, skills, interests, and sexual preferences [163].

Alexia Dini

2.7 Privacy policy analysis as a tool for user privacy awareness

Regulations, such as the GDPR, oblige service providers to inform the users about their practices regarding data collection and processing [280]. The existing method used for the portrayal of the rights and responsibilities of both the user and the service provider in terms of data collection, processing and sharing, are the privacy policies, which are legal texts that depict the practices that an organisation or company follows when handling the personal data of its users [233]. The introduction of the GDPR resulted in service providers having to adapt their privacy policies content to the new requirements, providing all the required information to the users.

The GDPR is a regulation which is aimed at data controllers, however the users are what

Table 2.5: Summary of related work for privacy policy analysis as a tool for user privacy awareness

Area	Cit.	Method
Websites	[70]	Information extraction techniques
Apps	[13]	Information extraction techniques
Websites	[318]	Machine Learning
General	[32]	Semantic analysis techniques
General	[120]	Machine Learning
Websites	[280]	Machine Learning and NLP techniques
General	[45]	NLP techniques
General	[125]	Deep learning
General	[51]	Machine Learning
GDPR Related	[181]	Machine Learning
	[202]	Semantic analysis techniques
	[123]	NLP
	[285]	Machine Learning and NLP techniques
	[177]	Machine Learning and NLP techniques
	[300]	Semantic analysis techniques

Alexia Dini

the content is really about. The aim of GDPR is to protect the users and their rights, which are recorded as the *Rights of the Data Subject* in Chapter 3 of the GDPR [93]. Furthermore, GDPR Articles 12-14 designate that data controllers *must communicate any mandatory information or information relating to data processing to the user in a concise, transparent, intelligible and easily accessible form, using clear and plain language, as well as information necessary to ensure a fair and transparent processing* [93].

In the last years, several studies have focused on the analysis of privacy policies in order to assist users by making the content of the privacy policies easier to understand or by automating their assessment [80, 280]. A solution that can be used to analyze the text of privacy policies of websites and display the personal information collected is presented in [70], where information extraction techniques are exploited to extract data collection practices. *PolicyLint* [13] is a tool that also uses information extraction methods in order to extract information from the privacy policies of 11,430 apps and to detect any existing policy contradictions in the policy content, but the dataset is not publicly available. While in [318], the privacy policies of websites are analyzed with the aim to assess what data practices are being used and described in the text of the policies that the users are presented with. In [32], the authors propose a representation of data practice descriptions in privacy policies as semantic frames in order to identify incompleteness and the different values attached to four categories of data actions, i.e. collection, retention, use, and transfer.

The use of Machine Learning techniques has also been explored in the literature for the analysis of privacy policies. In [120], Machine Learning models are being used to classify different segments of privacy policies with a view to examine the integrity of the content based on the current data practices. The *PrivacyGuide* tool is proposed in [280], which uses Machine Learning and Natural Language Processing techniques (NLP) for the analysis of 45 policies from the most accessed websites in Europe. Another tool is presented in [45], namely the *PIExtract dataset*, which uses NLP techniques such as named entity recognition, to automatically extract information from privacy policies and assist users to comprehend what personal information is collected about them and shared with third parties. An automated framework for privacy policy analysis, *Polisis*, is presented in [125], that uses a deep learning system enabling scalable, dynamic, and multi-dimensional queries on natural language privacy policies, displaying them to the users in a comprehensive manner. In [51], a system is proposed for the automatic extraction of fine-grained data practices from privacy policies and train models to predict the privacy policies of apps.

In relation to the GDPR, a number of works in the literature are engaged with the analysis of privacy policies against the GDPR with the aim to provide insightful outcomes for data subjects, i.e. the users. An approach for the automatic analysis of the content of privacy policies aiming to discover any violations against the GDPR Article 13 is proposed in [181], i.e. checking the inconsistencies between GDPR and privacy policies. The work in [202] uses semantic text-matching techniques to find the consistencies between privacy policies and the relevant GDPR articles. NLP techniques are used in [123] for the extraction of the data practices existing in privacy policies, while they also investigate the existence of mandatory information by encoding GDPR rules. In another work, [285], a tool is presented for automatically checking if the content of a privacy policy is complete conforming to the GDPR provisions. The authors use Machine Learning and NLP techniques for the automatic classification of the content of a given privacy policy. In [177] the authors analyse the text of privacy policies for the automated detection of GDPR violations in the tool they present *Claudette*, while in [300], *Complicy* is proposed, a tool for the evaluation of the GDPR alignment of privacy policies in the area of web platforms. Table 2.5 summarises the related work for privacy policy analysis as a tool for user privacy awareness presented in this section.

Alexia Dini

2.8 Machine Learning Techniques

The development and use of Machine Learning techniques have been associated with the advancements in Big Data technologies. Machine learning techniques aim to to solve problems based on historical examples. According to Arthur Samuel, a pioneer of artificial intelligence research, Machine Learning is a *“Field of study that gives computers the ability to learn without being explicitly programmed* [193]. Machine Learning can be used for data analytics, for extracting information from previous knowledge, for predictive modeling and applying the knowledge to predict new instances, and for decision-making.

Machine Learning algorithms are categorised into supervised and unsupervised algorithms. The difference between these categories is the presence of labels in the training dataset. Supervised machine learning engages the use of input attributes as well as the use of fixed target attributes. These algorithms pursue to predict and classify the target attribute, while their performance measures depend on the total number of the correctly predicted or classified target attribute [11]. The supervised learning algorithms are further classified into classification and regression algorithms. Unsupervised machine learning is concerned with pattern recognition without the engagement of a target attribute. In essence, this means that all the variables are used as inputs.

2.8.1 Supervised Machine Learning Algorithms

Supervised learning aims is to create a model of the distribution of class labels based on predictor features [157]. Supervised machine learning algorithms generate a function that maps inputs to required outputs. A set of training data containing labels is fed to the algorithm, which, based on the data provided, will learn a rule and uses it to predict the labels for new observations. Supervised machine learning algorithms are categorised into regression and classification algorithms, where regression-based methods aim to predict outputs based on input variables, while classification-based methods intend to identify the category that a set of data items belongs to.

A classification algorithm is an algorithm that uses a training dataset in order to learn and then assigns new data points to a specific class [253]. A classification task can be a binary or a multi-label classification task. In binary classification there are two possible outcomes, while multi-label classification tasks can have more than two possible outcomes. In the following paragraphs, we describe different supervised machine learning algorithms

Alexia Dini

that are used in classification.

K Nearest Neighbour:. This algorithm, also known as kNN, is a classification algorithm. It assigns the class of the nearest of a set of previously labelled points to an unlabelled sample point. kNN is not based on any underlying data distribution and is called non-parametric. For a given dataset, the algorithm predicts the relation between the unseen data and the existing data, and based on the prediction, it attributes the the new data to the predominant class that has the best match with it. The algorithm performance is based on the proper selection of the value of a variable parameter, known as k, which is the count of the nearest neighbours of the new data point. After that, the Euclidean distances of the predominant points in the data set from the new data point are calculated., which are used to find the category to which the majority of the nearest neighbours belong and perform the classification. The formula to calculate the Euclidean distance is shown in Equation 2.1 [169]. The Euclidean distance represents the shortest distance between two points and is calculated using the well-known Pythagorean theorem.

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{2.1}$$

Support Vector Machine:. The Support Vector Machine algorithm, or SVM, is an advanced supervised algorithm that can be used both for regression and classification problems, and it can deal with continuous and categorical instances. The SVM algorithm's goal is to create a best decision boundary that can separate n-dimensional space into classes by a clear margin widest possible, known as hyperplane, in order to put new data points in the correct category. Two types of SVM exist, Linear SVM, for linearly separable data, and Non-linear SVM, for non-linearly separated data. SVM shows a distinct increase in performance , when the " n of the n-dimensional space is greater than the total size of the sample set, therefore it is a good choice when dealing with high-dimensional data [253]. The SVM algorithm makes predictions based on Equation 2.2 [284], where K(x, x_i) is a kernel function that defines one basis function for each example in the training set. The target function of SVM pursues to minimise the error in the training dataset, while maximising the margin between the two classes, a mechanism that prevents overfitting.

$$y(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0 \tag{2.2}$$

Alexia Dini

Naïve Bayes:. The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification, based on applying the Bayes theorem from Bayesian statistics, with a strong assumption about the independence of the attributes given the class. One advantage of the algorithm is that it takes a small amount of training data for the estimation of the necessary parameters for classification, delivering great classification accuracy. There exist three models of the Naïve Bayes classifiers, the Multivariate Bernoulli model, the Multinomial model and the Probabilistic Model [149]. The Naïve Bayes classifier classifies a new instance by assigning the most probable target value v_{MAP}, given the attribute values a_1, a_2,..., a_n that describe the instance (Equation 2.3).

$$v_{MAP} = \arg\max P\left(C_j \mid a_1, a_2 \cdots a_n\right) \tag{2.3}$$

Based on the Bayes theorem this equation can be written as can be seen in Equation 2.4.

$$\begin{aligned} v_{MAP} &= \arg\max \frac{P\left(a_1, a_2 \cdots a_n \mid C_j\right) P\left(C_j\right)}{P\left(a_1, a_2 \cdots a_n\right)} \\ &= \arg\max P\left(a_1, a_2 \cdots a_n \mid C_j\right) P\left(C_j\right) \end{aligned} \tag{2.4}$$

The Naïve Bayes classifier is based on the assumption that the attribute values are conditionally independent given the target value, i.e. the assumption is that given the target value of the instance, the probability of observing the conjunction a_1, a_2, ..., a_n is the product of the probabilities for the individual attributes (Equation 2.5).

$$P\left(a_1, a_2 \cdots a_D \mid C_j\right) = \prod_i P\left(a_i \mid C_j\right) \tag{2.5}$$

Substituting this into Equation 2.4, we have the formula for the Naive Bayes classification algorithm(2.6) [136].

$$v_{NB} = \underset{C_j \in C}{\arg\max} P\left(C_j\right) \prod_i P\left(A_i \mid C_j\right) \tag{2.6}$$

Multi-label KNN:. Multi-label KNN or MLKNN is a multi-label lazy learning algorithm, adapted from the traditional KNN algorithm for multi-label data. For each unseen instance, its K nearest neighbours in the training set are identified and based on statistical information gained from the label sets of these neighbouring instances, the Maximum a Posteriori (MAP) principle is utilised to determine the label set for the unseen instance [338]. Given an instance x and its associated label set $Y \subseteq \mathscr{Y}$, it is assumed that KNNs are considered in the MLKNN method. Let $\vec{y}_x$ be the category vector for x, where its lth component $\vec{y}_x(l)(l \in \mathscr{Y})$ takes

Alexia Dini

the value of 1 if $l \in Y$ and 0 otherwise. In addition, let $N(x)$ denote the set of KNNs of x identified in the training set. Thus, based on the label sets of these neighbours, a membership counting vector can be defined as [339]:

$$\vec{C}_x(l) = \sum_{a \in N(x)} \vec{y}_a(l), \quad l \in \mathscr{Y} \tag{2.7}$$

where $\vec{C}_x(l)$ counts the number of neighbours of x belonging to the lth class. For each test instance t, MLKNN firstly identifies its KNNs $N(t)$ in the training set. Let H_1^l be the event that t has label l, while H_0^l be the event that t has not label l. Furthermore, let $E_j^l(j \in \{0, 1, \ldots, K\})$ denote the event that, among the KNNs of t, there are exactly j instances which have label l. Therefore, based on the membership counting vector $\vec{C}_t$, the category vector $\vec{y}_t$ is determined using the following MAP principle [339]:

$$\vec{y}_t(l) = \underset{b \in \{0,1\}}{\arg\max} P\left(H_b^l \mid E^l_{\vec{C}_t(l)}\right), \quad l \in \mathscr{Y} \tag{2.8}$$

Using the Bayesian rule, Equation 2.8 can be rewritten as

$$\vec{y}_t(l) = \underset{b \in \{0,1\}}{\arg\max} \frac{P\left(H_b^l\right) P\left(E^l_{\vec{b}_t^l(l)} \mid H_b^l\right)}{P\left(E^l_{\vec{\sigma}_t(l)}\right)} = \underset{b \in \{0,1\}}{\arg\max} P\left(H_b^l\right) P\left(E^l_{\vec{C}_t(l)} \mid H_b^l\right). \tag{2.9}$$

As shown in Equation 2.9, in order to determine the category vector $\vec{y}_t$, all the information needed is the prior probabilities $P\left(H_b^l\right)(l \in \mathscr{Y}, b \in \{0, 1\})$ and the posterior probabilities $P\left(E_j^l \mid H_b^l\right)(j \in \{0, 1, \ldots, K\})$. Actually, these prior and posterior probabilities can all be directly estimated from the training set based on frequency counting [339].

Binary Relevance:. The Binary Relevance (BR) algorithm is a typical and efficient problem transformation approach used in multi-label classification. The BR algorithm transforms a multi-label problem by breaking it down into q independent binary classification problems in order to perform single-label learning. Each binary classification problem then relates to one class label in the label space $\mathcal{Y}$) [337]. Each binary classifier is trained separately for each label and the final classification result is the combination of the results of all the binary classifiers. For each class label λ_j determines a binary training set of $\mathcal{D}_j$ from the initial training set as shown in Equation 2.10, while the pseudo-code for the algorithm can be found at Algorithm 1 [337]. Let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional instance space and let $\mathcal{Y} = \left\{\lambda_1, \lambda_2, \ldots, \lambda_q\right\}$ denote the label space, consisting of q class labels. The goal

Alexia Dini

of multi-label learning is to induce a multi-label predictor $f : \mathcal{X} \mapsto 2^{\mathcal{y}}$ from the multi-label training set $\mathcal{D} = \left\{\left(\boldsymbol{x}^i, \boldsymbol{y}^i\right) \mid 1 \leqslant i \leqslant m\right\}$. Here, for each multilabel training example $\left(\boldsymbol{x}^i, \boldsymbol{y}^i\right), \boldsymbol{x}^i \in \mathcal{X}$ is a d-dimensional feature vector $\left[x_1^i, x_2^i, \ldots, x_d^i\right]^\top$ and $\boldsymbol{y}^i \in \{-1, +1\}^q$ is a q-bit binary vector $\left[y_1^i, y_2^i, \ldots, y_q^i\right]^\top$, with $y_j^i = +1(-1)$ indicating that y_j^i is a relevant (or irrelevant) label for x^i.[2] Equivalently, the set of relevant labels $Y^i \subseteq \boldsymbol{y}$ for $\boldsymbol{x}^i$ corresponds to $Y^i = \left\{\lambda_j \mid y_j^i = +1, 1 \leqslant j \leqslant q\right\}$. Given an unseen instance $x^* \in \mathcal{X}$, its relevant label set Y^* is predicted as $Y^* = f(x^*) \subseteq \dagger$.

$$\mathcal{D}_j = \left\{\left(\boldsymbol{x}^i, y_j^i\right) \mid 1 \leq i \leq m\right\} \tag{2.10}$$

Algorithm 1 Pseudo-code of BinaryRelevance Algorithm

```
Data: $\mathcal{D}$, $\mathcal{B}$, $x^*$
$\mathcal{D}$: the multi-label training set $\{(x^i, y^i) \mid 1 \leq \mathrm{i} \leq \mathrm{m}\}$
  ($x^i \in \mathcal{X}$, $y^i \in \{-1, +1\}^q$, $\mathcal{X} = \mathcal{R}^d$, $\mathcal{Y} = \{\lambda_1, \lambda_2, ..., \lambda_q\}$)
  $\mathcal{B}$: the binary learning algorithm
  $x^*$: the unseen instance ($x^* \in \mathcal{X}$)
Result: $Y^*$: the predicted label set for $x^*$ ( $Y^* \subseteq \mathcal{Y}$)

Process:
for $j = 1$ to $q$ do
    Derive the binary training set $\mathcal{D}$ according to Eq. 2.10
    Induce the binary classifier $g_j$: $\mapsfrom \mathcal{B}(\mathcal{D}_j)$
end
return $Y^* = \{\lambda_j \mid g_j(x^*) > 0, 1 \leq \mathrm{j} \leq \mathrm{q}\}$
```

BRkNN:. The BRkNN classifier is an adaptation of the kNN algorithm for multi-label classification. Theoretically, the algorithm is equivalent to using the single-label kNN method combined with a Binary Relevance setup, but runs much faster [269]. BRkNN extends the kNN algorithm by making independent predictions for each label through a single search of the k nearest neighbours. This assists in the algorithm being faster than a combination of BR with kNN, and can be particularly useful for problems with large sets of labels and requirements for low response times. There are two extensions to the BRkNN algorithm, BRkNN-a and BRkNN-b, both of which are based on the calculation of confidence scores for each label from BRkNN. The BRkNN-a classifier returns the labels that give the highest score even if these labels are lower than the threshold, as in most multi-label datasets it is

Alexia Dini

not common to have an empty set of labels. The BRkNN-b classifier reduces the cardinality of the labels between the predicted and the actual label sets.

ClassifierChains:. The Classifier Chains model (CC) is based on the BR method as it engages with the construction of L binary classifiers, where every classifier's task is the prediction of the relevance of one label. The attribute space for each binary classifier is extended with the 0/1 relevance of all previous classifiers, forming a classifier chain [334]. The training procedure of the CC algorithm is laid out in Algorithm 2 [238, 239]. Suppose the training example (x,S), where S $\subseteq$ L is represented by the binary feature vector $((l_1, l_2, \cdots, l_{|L|}) \in \{0, 1\}^{|L|}$, and x is an instance feature vector.Therefore, a chain $C_1, \cdots, C_{|L|}$ of binary classifiers is formed. Each classifier C_j in the chain has the task to learn and predict the binary association of label l_j given the feature space, expanded by all previous binary relevance predictions in the chain $l_1, \cdots, l_{j-1}$. Classification starts at C_1 and reproduces along the chain: C_1 determines $\Pr(l_1 \mid x)$ and every subsequent classifier $C_2 \cdots C_{|L|}$ predicts $\Pr\left(l_j \mid x_i, l_1, \ldots, l_{j-1}\right)$ [238]. This classification process is portrayed in Algorithm 3 [238, 239].

Algorithm 2 Pseudo-code of CC training procedure for training set $\mathcal{D}$ and label set $\mathcal{L}$ of L labels

```
TRAINING (D = {($x_1$, $S_1$),..., ($x_N$, $S_N$)})
for $j \in 1...|L|$ do
    ▷single-label transformation and training
    $D_j^{'} \leftarrow \{\}$
    for (x,S)∈ D do
        $D^{'} \leftarrow D^{'} \cup ((x^{'}, l_1, ..., l_{j-1}), l_j)$
    end
    ▷ train $C_j$ to predict binary relevance of $l_j$
    $C_j : D^{'} \rightarrow l_j \in \{0,1\}$
end
```

2.8.2 Unsupervised Machine Learning Algorithms

Unsupervised machine learning is also known as clustering analysis. The major difference between unsupervised and supervised machine learning machine learning is that in unsupervised machine learning there is no training data set, and as such no cross-validation is

Alexia Dini

Algorithm 3 Pseudo-code of CC prediction phase for a test instance x

```
CLASSIFY(x)   y ← { }
for $j \leftarrow 1$ to $|L|$ do
    Y ←Y ∪$\big(l_j \leftarrow C_j : \big(x, l_1, \cdots, l_{j-1}\big)\big)$
end
$(x, Y)$▹ the classified example
```

required. As no human interference is required, unsupervised machine learning is a data-driven process, mainly used for the identification of trends, exploratory purposes, etc. The most prevalent unsupervised learning tasks are clustering, feature learning, dimensionality reduction, anomaly detection, etc. [246]. In the paragraphs that follow, we present different unsupervised machine learning algorithms.

K-Means:. The K-means algorithms is one of the simplest unsupervised learning algorithms, that is also fast and powerful, providing reliable results. The algorithm assigns the data points to a cluster in a manner that the amount of the squared distance between the data points and the centroid is the smallest possible [246]. The pseudo-code for the K-means algorithm is depicted in Algorithm 4 [333]. The algorithm creates k points as initial centroids on the spot, where k is a paremeter specified by the user. Then each point is assigned to the cluster with the closest centroid, and the centroid of each cluster is updated using the mean of the data points of each cluster. As some data points may move from one cluster to another, new centroids are calculated and the data points are assigned to the suitable clusters. This process is repeated, until no points move to other clusters and the centroids remain the same. The Euclidean distance is used in this algorithm to find the distance between data points and centroids [333].

DBSCAN:. Density-based spatial clustering of applications with noise (DBSCAN) is an algorithm for density-based clustering, a technique used to separate high-density clusters from low-density clusters that are used in model building. In DBSCAN, given a set of points in some space, the algorithm groups together points that are closely packed together, while it can find clusters of multiple shapes and sizes in a big amount of data that is both noisy and that contains outliers [246]. The pseudo-code for the steps of the DBSCAN can be seen in Algorithm 5 [147]. The DBSCAN algorithm is a density based algorithm that determines clusters with random shape, using two parameters: the radius of the cluster (Eps) and the

Alexia Dini

Algorithm 4 Pseudo-code of the K-means algorithm

```
Require: D = {$d_1$, $d_2$, $d_3$, ..., $d_i$, ..., $d_n$} set of n data points
Require: k: Number of clusters
Ensure: A set of $k$ clusters
Steps: Arbitrarily choose k data points from D as initial centroids
repeat
    Assign each point $d_i$ to the cluster which has the closest centroid
    Calculate the new mean for each cluster
until                                                          ▷;
Convergence criteria is met.
```

minimum required points inside the cluster (MinPts).

Algorithm 5 Pseudo-code of the DBSCAN algorithm

```
Function DBSCAN (Dataset D, Eps, MNinPts)
1: Select an arbitrary object $P$ in $D$;
2: Retrieve all objects density-reachable from $P$ by arbitrary/random Eps and MinPts values;
3: if $P$ is a core object then a cluster is formed;
4: if $P$ is a border object then no objects are density reachable from P and DBSCAN visits the next
object of the dataset;
5: else assign $P$ to noise object;
6: Continue the process (from Step 1) until all of the objects have been processed.
end
```

Alexia Dini

Agglomerative Hierarchical Clustering:. Agglomerative Hierarchical Clustering is a type of hierarchical clustering used to group objects in clusters according to their similarity. A bottom-up approach is being used in this algorithm, where each object is first treated as an individual cluster (leaf). At each step of the algorithm, two clusters that are considered to be the most similar are merged into a single large cluster, called node. The procedure is repeated until all points belong to one single cluster, the root [148]. The result is a dendrogram, which is a tree-based depiction of the elements [246]. This procedure is depicted in Figure 2.1 [148].

Mean-shift clustering:. Mean-shift clustering is a clustering algorithm that does not require prior knowledge of the number of clusters or constraints on cluster shape. The aim of

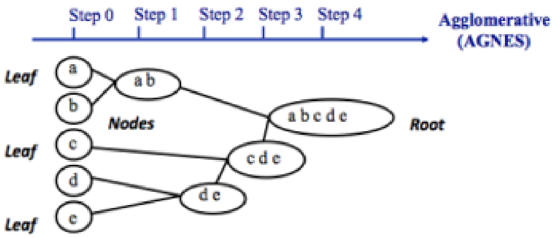


Figure 2.1: Agglomerative Hierarchical Clustering procedure [148]

the algorithm is to find “blobs" in a smooth distribution or density of samples. The algorithm works by updating centroid candidates to be the mean of the points in a given region, by filtering the candidates to remove duplicates. It is a computationally expensive method that does not work well in cases of high dimension [246]. The flowchart of the Mean-shift clustering algorithm is depicted in Figure 2.2 [182].

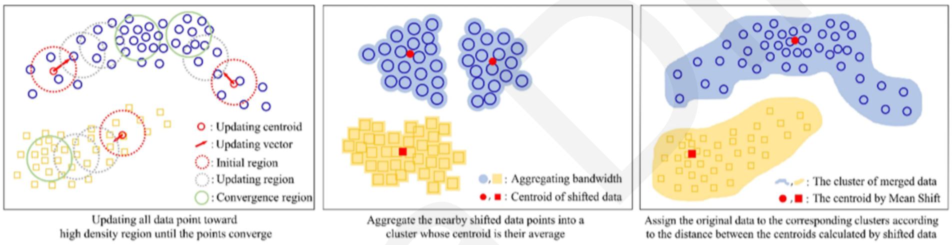

○ : Updating centroid
: Updating vector
: Initial region
: Updating region
: Convergence region
Updating all data point toward
high density region until the points converge
●,■ : Aggregating bandwidth
●,■ : Centroid of shifted data
Aggregate the nearby shifted data points into a
cluster whose centroid is their average
: The cluster of merged data
●,■ : The centroid by Mean Shift
Assign the original data to the corresponding clusters according
to the distance between the centroids calculated by shifted data

Figure 2.2: The flowchart of the Mean-shift clustering algorithm [182]

2.9 Discussion and Conclusion

The literature review and background work suggest that smart home devices and fitness trackers have the potential to greatly benefit users in terms of convenience, energy efficiency, and health and wellness. However, user awareness about data collection and sharing practices of the service providers, as well as the possibility of inference risks, are areas that require further research especially after the introduction of the GDPR. User awareness of the risks associated with the use of smart home technology and fitness trackers is crucial in ensuring that their personal information is protected.

The systematic quantitative literature review (SLR) we conducted in order to explore the state-of-the-art in the field of user privacy protection in IoT showed that research in this area is at early stages [276]. As the GDPR handles personal data in every sector that it applies and is involved with the rights of the users whose personal data are being processed through the introduction of the eight user rights that all EU citizens are entitled to regarding their

Alexia Dini

data, the aim of the GDPR is to give control of personal data to users. In order to address the privacy challenges that have been identified through the review in this Chapter, user-centric solutions are essential, such as a GDPR-compliant privacy framework that can protect personal data in the IoT and empower users with greater control over their personal data, which we discuss in Chapter 3. What makes this work different from prior research in the area is that the proposed framework is based on a number of steps and processes that provide users with practical functionalities and tools to manage their personal data generated by IoT devices. This approach goes beyond theoretical concepts by offering tangible mechanisms for users to exercise their rights, make informed decisions, and protect their privacy effectively. Furthermore, our research recognises the unique challenges and requirements posed by the GDPR in the context of IoT. The proposed framework aligns with the GDPR principles and provisions, addressing the complexities of data collection, sharing, and privacy risks associated with IoT devices. By tailoring the framework to this specific legal framework and technological landscape, a targeted and practical solution for GDPR-compliant privacy protection in IoT is provided.

In relation to privacy inference risks in IoT, most of the approaches found in the literature utilise the data collected from smart devices in order to provide better services, such as elderly monitoring, improvement of smart home applications, health-care or security [39, 310], without taking into consideration the protection of the user privacy when handling the users data. The diversity of the work presented in this thesis is that smart home devices and fitness trackers data are utilised in order to explore how the user privacy can be compromised and inform the users about any inferences that can be drawn from their data. Furthermore, the limitations of some of the approaches discussed in Sections 2.5.3 and 2.5.4 are that even though they show that a number of inferences are possible from fitness trackers and smart home devices data that pose a threat to the users' privacy, none of these works aims to notify the users about them and raise user awareness, and this is what makes the work presented in this thesis different from them, as in this thesis the aim is to inform the users about any data privacy vulnerabilities that are identified through a dedicated web application, increasing their awareness. While Psychoula et al. [230] were occupied with user privacy awareness in the area of wearables and IoT services by presenting a framework that could be used as guidance to developers and service providers in order to integrate privacy risk user awareness in their products, no other work to the best of our knowledge has been involved with raising user awareness in relation to the inferences that can be extracted about the users from their smart home devices and fitness trackers data.

Alexia Dini

Furthermore, this doctoral thesis aims to further increase the awareness of users by examining the privacy policies of smart devices introducing the SpotAware approach. To the best of our knowledge, no existing work focuses on the automatic classification of privacy policy text extracting information regarding the eight GDPR user rights present and the data inferences that can be made about the users based on the collected data as described in the policy text.

In Section 2.8, the concept of Machine Learning was introduced. In the PrivacyEnhAction web application for the inference detection from smart home devices, we have adopted and implemented the K-Means algorithm, in order to analyse motion sensor and smart water meter data and identify patterns or clusters based on motion, water usage, or combined data patterns. The use of this particular algorithm assisted in grouping similar sensor readings together and understanding different usage or motion profiles. The choice of this algorithm was made after experimenting with different algorithms and considering the nature of the available data, as it was found to be the best fit for the research objectives and data characteristics. Furthermore, the SpotAware approach was implemented in the PrivacyEnhAction web application by adopting the BinaryRelevance multi-label classifier. The choice for this algorithm was made after conducting experiments with a number of multi-label classifiers and evaluating the results, and also due to the fact that it does not assume any label dependencies, making it suitable for our research objectives where the labels were independent.

Alexia Dini

Chapter 3

A user-centric privacy framework for personal data protection in IoT

Alexia Dini

As the Internet of Things continues to expand, the collection and use of personal data have become increasingly ubiquitous. From smart homes to wearable fitness trackers, IoT devices have the ability to collect and transmit vast amounts of personal data in real-time. While these technologies offer many benefits to users, such as convenience and improved health outcomes, they also raise concerns about user data and privacy protection. To address these concerns, a user-centric privacy framework is essential for protecting personal data in the IoT. Such a framework should take into account the privacy preferences of the users, as well as the various types of IoT devices and services they use. The framework should provide users with greater control over their personal data, while also ensuring that companies and organisations are accountable for protecting these data.

This chapter introduces the concept of a user-centric privacy framework for personal data protection in IoT. By reviewing the state-of-the-art literature we first define a number of characteristics that such a framework should possess in order to empower the users to be in control of their data collected by their IoT devices. Then the architecture of the framework is presented, along with a description of the proposed steps that constitute the processes of the framework demonstrating how the users can be provided with the functionalities and the tools needed in order to be in control of their personal data created by IoT devices. We also discuss the findings that were observed through this work regarding the most commonly used techniques in the literature for addressing the proposed characteristics, and we define the context on which the rest of this doctoral thesis is based. The work presented in this chapter is based on our research published in the Elsevier *Internet of Things; Engineering Cyber Physical Human Systems* journal [160].

3.1 Background

The enormous amount of data collected and shared among IoT resources has raised an important issue regarding the privacy and the awareness of the users about how their data are collected and shared. For the typical user, it is very difficult to control the data shared by the devices she owns. Furthermore, a privacy risk occurs when data collection and processing leads to the leaking of personal information [286], known as an inference attack, where personal information of a user can be assumed by exploiting the data that the user has shared, making the identification of a user possible.

The IoT industry is now facing the GDPR [92], which was put in place in 2018. The regulation provides fundamental directions in order to accomplish an equitable treatment of the third parties and the users. Additionally, it radically changes how data is handled in every area applied and creates standards for the protection of user data in IoT. The GDPR addresses issues, such as the kind of data that can be treated and under which circumstances, the purpose for collecting this data, the amount of data that can be collected, the retention period of the data, and the information that the users should know about their collected data. Therefore, it is involved more with the rights of the users whose data are being processed, which are: *the right to be informed, the right of access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object and the right to avoid automated decision-making* [65]. In simple words, the GDPR aims to give control of personal data back to the user. Since the IoT depends upon effusive user data collection and sharing, the probability of risks for the user privacy increases. The GDPR calls for “increased user involvement in protecting their data by enabling them to control what is collected about them, when, by whom and for what purposes" [26]. In this respect, traditional privacy approaches must advance from focusing on the service providers to the users. In order to address these issues, a user-centric privacy framework is essential for protecting personal data in the IoT, which should provide users with greater control over their personal data. This is why we define the first Research Question that this thesis aims to answer: **RQ1:*“What are the characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess?"***. The term “Privacy Framework" refers to any model or solution with the aim to provide a structure for the management of personal data that can be used by developers, organisations, services, etc., in order to comply with the GDPR.

Reviewing the recent literature regarding the protection of the user privacy in IoT has shown that research in this area is at early stages [276]. A study was carried out by Wachter [308],

Alexia Dini

where the author investigates the tension between data privacy and identifiability in the domain of IoT, by looking into earlier academic and policy discussions, in order to understand how the GDPR principles can safeguard the user privacy in IoT. Her work focuses on four user privacy challenges in GDPR, which were identified through her review: (1) *profiling, inference, and discrimination*; (2) *control and context-sensitive sharing of identity*; (3) *consent and uncertainty*; and (4) *honesty, trust, and transparency*. The motivation for this work originates from Wachter's analysis. Using the identified challenges, we devised a number of characteristics that a GDPR compliant privacy framework in IoT should possess, providing an answer to **RQ1**.

With a view to understand how Wachter's challenges can be addressed in the scope of the GDPR, it is important to designate the techniques and methods that can be used to protect the users' personal data and privacy in the IoT domain. Wachter performed an analysis of the state-of-the-art literature in the scope of the GDPR in order to assess how the directions of the GDPR can assist in addressing the user privacy issues in IoT. Based on this analysis and by thoroughly examining the requirements and risks of each challenge, we have identified a list of characteristics that an IoT GDPR-compliant privacy framework should possess in order to empower the users to be in control of their personal data and privacy.

We have also performed a mapping of user-centric privacy preserving approaches from the state-of-the-art literature to these characteristics, which are classified under each challenge, in order to analyse how each characteristic is addressed for user privacy protection and to get an insight regarding the methods and techniques used for the protection of user personal data and privacy in various domains of IoT, such as smart homes, smart buildings, smart grids, etc. This allows us to contribute to the research by providing a basis for the design and development of effective user privacy frameworks in IoT, that can be used by researchers for carrying out further research in the area, or by practitioners who can incorporate the characteristics to their platforms or systems for better protection of their users.

3.2 The GDPR Challenges

Wachter's work was based on major areas of concern in regulating the IoT, as identified by Peppet in [224], namely "*discrimination, privacy, security, and consent*". Even though Peppet's work considers IoT challenges under the existing American policies context, Wachter's thematic review was inspired by his analysis and the author adopted and applied it to the European policy context. Wachter has proposed four challenges related to user privacy follow-

Alexia Dini

ing the taxonomy of Peppet, which describe the outcomes emerging from issues of privacy in IoT environments under GDPR. The author examines how these challenges can be addressed, considering "*the relevant GDPR standards for transparency, data storage, access, rectification, and deletion, informed consent, notification duties, automated decision-making and profiling, and privacy by design and by default*" [308].

The first challenge, "*Profiling, inference and discrimination*" (CH1), rises from the potentiality that user data, such as user identity, can be connected with other data created by IoT devices. Such a scenario can result to user profiling, data inferences and discrimination against the user. Many methods exist that can be used for profiling, for example when collecting data that can be used to draw conclusions about a user, or when using and combining data sets generated by IoT devices with other third parties. An pertinent scenario can be observed with fitness trackers, where privacy vulnerabilities form a major privacy issue, as the users wear these devices almost steadily [258], making the leakage of data a significant threat. Additionally, the combination of data regarding the user's physical state, such as the heart rate, with the user's movements, make it possible for inferences to occur. Profiling can be used to obtain a user's political or religion beliefs, sexual preferences or health status, leading to privacy breaches [252]. Inference techniques are used in order to acquire additional personal information from the available data collected from IoT devices. For example, step-based data collected from a wearable device may enable the determination of the user's location [323] from attackers, or the monitoring of the exercising routines of a user [8]. Furthermore, discrimination may be possible from the exploitation of medical user data, such as psychiatric behaviour or HIV suffering, from employers or health insurance companies [91].

The second challenge, "*Control and context-sensitive sharing of identity*" (CH2), has been identified by the possibility that the users' sensitive information may be revealed to others, as the users are not always able to define constraints for their personal data and they lack control over it. IoT services for data sharing are available to users by subscribing to those services, who are usually untrusted entities aiming to get access to user data in order to gain financial benefits or having other motives. In this sense, data privacy must be addressed by providing control to the users for the sharing of their data [321]. For example, the health-related Internet of Things (H-IoT) provides real-time monitoring of patients; however, the patients have no real control of their personal data generated and analysed by the H-IoT other than accessing it, putting them in a vulnerable position against undesirable exploitation [198]. For the purposes of this work, we do not address only the issue of identity sharing but the sharing of any personal information.

Alexia Dini

Wachter has identified the third challenge, "*Consent and uncertainty*" (CH3), by the uncertainty of data generated by IoT devices. Users are surrounded by many such devices that collect and process their data and send them to third parties. The users do not always have the means to define their privacy preferences for the sharing and processing of their personal data, in order to provide their informed consent. When the users are uncertain of whether their choice of privacy settings will enable data inferences to be made about them, the need for informed consent is not satisfied [306]. One common example of the lack of the provision of user informed consent is the tracking of visitors in a shopping mall, where a monitoring system collects the MAC addresses of WiFi or Bluetooth connected devices, while the visitors are not aware that their data are being collected [201]. In another case, when the first IoT botnet was discovered in 2013, it became evident that the botnet - which consisted of IoT devices, such as smart TVs, baby monitors, etc. - was collecting user personal information, like user names and telephone numbers, while at the same time monitoring user activities without their consent [327]. In both examples, the importance of informed consent is emphasised calling for the need to make the users understand how their data can be exploited. In another recent example, Vizio, an electronic product development company, was fined 2.2 million dollars for selling 11 million Smart TVs with a specific software installed to track the viewing habits of unaware users [254].

The fourth challenge, "*Honesty, trust, and transparency*" (CH4) has been identified by the user limitation in the supervision and transparency in the management of collected data in the IoT domain. This makes privacy breaches easy to occur, weakening the trust between users and third parties. For this reason, trust relationships must be established between devices, users and third parties for data sharing, where appropriate permissions can be set. User trust and confidentiality can be enhanced by the provision of authentication and access control mechanisms [306]. Additionally, the user must be empowered to know who has access to her data, what data has been shared and how it will be used [104], as this information is crucial for user trust and acceptance. The Internet of Toys (IoToys), which is part of the IoT domain, is a very good example for the importance of transparency and trust in IoT. Connected toys manufacturers store data such as children's conversations with a smart doll or robot in the cloud. while their privacy policies briefly mention that the data will be used for purposes such as those surrounding "Services provided by our Trusted Partners" and "Advertising" [111]. These methods may violate users' privacy and without transparency, parents have no control over their children's data.

Alexia Dini

3.3 Characteristics and mapping to challenges

In Wachter's work, each challenge is analysed in relevance to how the GDPR addresses the issue, taking into consideration existing problems in the protection of user and data privacy. Wachter has benchmarked state-of-the-art literature against the GDPR in order to assess how the directions of GDPR can assist in resolving the user privacy issues in the IoT. This assessment has been proved to be helpful for the identification of the characteristics in this work. While the challenges identified by Wachter highlight the existing problems and issues that need to be addressed in IoT in the context of GDPR, the characteristics define the specific attributes that a user privacy-preserving framework in IoT should possess to effectively tackle those challenges. In simple words, the challenges represent the problems or areas of concern, while the characteristics represent the desired features or capabilities of a framework to mitigate those challenges. By thoroughly reviewing Wachter's analysis and using them as a reference point in relation to the GDPR, we were able to extract a comprehensive list of GDPR characteristics that are essential for an effective user-centric privacy framework in the IoT. These characteristics can serve as a guideline for designing and evaluating such frameworks.

Inference risks originating from data collection and sharing with third parties, where data can be combined with other data sets to obtain further information, are possible. This has led to the definition of the first characteristic, CR1, *Prevent inference*, which has been mapped to CH1. It is concerned with actively minimising or mitigating the potential for drawing sensitive conclusions or insights from user data. The goal is to prevent unintended disclosures or privacy breaches by implementing measures and techniques that limit the possibilities of inferring sensitive information from the collected data.

The second characteristic, CR2, *Provide data transformation*, has been identified from the possibility of failure in data protection using anonymization techniques leading to user tracking or enabling the linking to other data sets. This issue is recognised in GDPR, where data transformation techniques are required to protect user privacy. CR2 is also relevant to CH3, where issues, such as user consent and notice, need to be addressed using transformation techniques.

In order for transparency to be applicable in IoT, users need to be aware that their data are being collected and how they are processed. This has led to the definition of the third characteristic, CR3, *Provide user awareness on data collection*, which is mapped to CH4. This characteristic focuses specifically on ensuring that users are fully aware of the data that are

Alexia Dini

being collected from them, emphasising the need to inform and educate users about the types of data being collected, the purpose of the data collection, and the potential implications or risks associated with it. This characteristic aims to empower users by providing them with a clear understanding of what data are being collected about them, which enables them to make informed decisions regarding their privacy.

One of the main goals of GDPR is to provide users with control over the disclosure of their personal data. This has been identified as the fourth characteristic, CR4, *Provide control of personal data to users*, mapped to CH2 and CH3. Under CH2, it is important that users have full control of the sharing of their data, while under CH3, CR4 is viewed under the informed consent requirement. This characteristic emphasises the importance of empowering individuals with the ability to manage their personal data, recognising that users should have control over how their data are collected, used, and shared in IoT environments. It involves providing users with clear and understandable options to make informed choices about data collection and processing, including mechanisms for obtaining user consent and allowing them to specify their privacy preferences. Furthermore, providing control to users involves giving them the ability to modify or revoke their consent and preferences over time. Users should have the flexibility to change their privacy settings, limit data sharing, or even delete their data if desired. In the same manner, we have identified the fifth characteristic, CR5, *Provide monitoring and control of devices that collect data*, which maps to CH3, as it concerns user consent. This characteristic focuses on enabling users to have oversight and control over the behaviour and actions of IoT devices that collect their data, recognising the need for users to be aware of and have the ability to monitor the activities of the devices they interact with and emphasising the importance of transparency and accountability in the data collection process.

The sixth characteristic, CR6, *Provide tools for data management to users*, is extracted from the common need of CH2 to CH4 to provide users with appropriate tools that allow them to control the usage of their data (CH2), oversee and control how they generate and share data (CH3), and provide transparency (CH4). GDPR has rendered the need for data erasure under Article 17 (Right to erasure), leading to the seventh characteristic, CR7: *Provide ability for data erasure*. This complements CR6 and its mapping to CH2 to provide users with the means to erase or rectify their data. Transparency is also one of the key requirements of GDPR (Article 12). Following Wachter's analysis, this has been defined as the eighth characteristic, CR8, *Provide transparency*, mapped to CH2, where users should be given control and be aware of how their data are processed, and to CH4, where trans-

Alexia Dini

parency is vital for increasing user's trust in an IoT system. This characteristic encompasses a broader concept related to the overall transparency of the privacy framework, extending to transparency in terms of the framework's policies, practices, and procedures. It involves providing users with detailed information about the privacy practices, including data retention policies, security measures, third-party sharing, and any relevant privacy settings or controls. While there is a similarity between this characteristic and CR3, *Provide user awareness on data collection*, in that both characteristics aim to promote transparency and user empowerment, they differ in their specific focus. CR3 places emphasis on the user awareness and understanding of the specific data being collected, its purpose, and associated risks, ensuring that users have a clear picture of what information are being collected from them, while CR8 takes a broader view, encompassing all aspects of the privacy framework and ensuring openness and clarity in how data are managed and processed.

In Wachter's analysis it becomes clear that there is a mismatch on how the interests of the user are balanced against those of the third party, leading to the definition of the ninth characteristic, CR9, *Provide balance between users and third parties*, which maps to CH2.

The need for the enforcement of user privacy preferences in order to protect privacy has led to the definition of the tenth characteristic, CR10, *Provide enforcement of user privacy preferences*, mapped to CH3. This characteristic focuses on implementing mechanisms to ensure that user privacy preferences are effectively enforced. It involves developing technical and organisational measures to enforce the specified privacy preferences, such as access controls, data anonymization, or secure data storage. This characteristic aims to provide users with reassurance that their privacy choices are respected and implemented throughout the data life-cycle.

Under the same challenge, the general requirement for implementing privacy by design and by default for privacy protection is analysed by Wachter, leading to the eleventh characteristic, CR11, *Provide privacy by design or privacy by default*. Furthermore, the GDPR requirement for informed consent (Article 7) has assisted in the definition of the twelfth characteristic, CR12, *Provide ability to users to make informed consent decisions*. In order for the users to be able to provide their informed consent, they must be informed of the possible risks. This has led to the definition of the thirteenth and fourteenth characteristics, CR13 and CR14, *Estimate privacy risks of data collection/inference to users* and *Communicate risks of data collection/inference to users*, respectively, which have not been combined because they address different aspects of managing privacy risks in data collection and inference. CR13 focuses on the analysis and evaluation of potential privacy risks that can arise from

Alexia Dini

the collection of user data. It involves conducting assessments and evaluations to identify and quantify the risks associated with data collection and the potential inferences that can be drawn from that data. This characteristic aims to provide an understanding of the risks involved, allowing for informed decision-making and the implementation of appropriate mitigation strategies. CR14 emphasises the importance of effectively delivering these identified risks to the users. It involves developing clear and understandable communication channels and methods to inform users about the potential privacy risks associated with data collection and inference. This characteristic aims to enhance user awareness and understanding, enabling individuals to make informed choices and take necessary precautions to protect their privacy. The first characteristic, *Prevent inference* is distinct from characteristics CR13 and CR14, as they serve different purposes within the privacy framework. By having CR1 as a separate characteristic we recognises the importance of proactive privacy protection. Risk estimation and communication alone may not be sufficient to ensure user privacy, therefore we emphasise the need for organisations and service providers to take proactive steps to minimise the potential for drawing sensitive inferences from user data.

Additionally, in order to enhance user privacy in IoT systems, the users must be able to express their preferences regarding privacy, specifying the fifteenth characteristic, CR15, *Provide ability to users to specify their privacy preferences*. This characteristic focuses on empowering users by allowing them to actively participate in the decision-making process regarding their privacy. It involves providing mechanisms and tools through which users can define their privacy preferences, such as specifying the types of data they are willing to share or the purposes for which their data can be used. This characteristic aims to give users a sense of autonomy and control over their personal information.

The need for the prevention of excessive data collection according to GDPR's Article 7 is very important, since data minimization (Articles 5, 7) requires that data processing should only use as much data as needed to successfully accomplish a given task, leading to the last characteristic, CR16, *Prevent excessive data collection*, mapped to CH3. This characteristic is about limiting the amount of data collected to only what is necessary for the intended purposes. It involves implementing measures and mechanisms to ensure that data collection practices adhere to the principle of data minimization. By minimising data collection, unnecessary or excessive gathering of personal information is avoided, reducing the potential risks associated with data breaches or unauthorised use.

Even though the proposed characteristics have been defined for IoT, they are applicable in other areas as well, as similar issues have been identified and addressed in areas such as

Alexia Dini

Table 3.1: Mapping of GDPR characteristics to challenges

No	CHARACTERISTIC	CH1	CH2	CH3	CH4
CR1	Prevent inference	✓			
CR2	Provide data transformation	✓		✓	
CR3	Provide user awareness on data collection				✓
CR4	Provide control of personal data to users		✓	✓	
CR5	Provide monitoring and control of devices that collect data			✓	
CR6	Provide tools for data management to users		✓	✓	✓
CR7	Provide ability for data erasure		✓		
CR8	Provide transparency		✓		✓
CR9	Provide balance of privacy between users and third parties		✓		
CR10	Provide enforcement of user privacy preferences			✓	
CR11	Provide privacy by design or privacy by default			✓	
CR12	Provide ability to users to make informed consent choices			✓	
CR13	Estimate privacy risks of data collection/inference to users			✓	
CR14	Communicate risks of data collection/inference to users			✓	
CR15	Provide ability to users to specify their privacy preferences			✓	
CR16	Prevent excessive data collection			✓	

✓= subject addressed; (blank) = not addressed

Alexia Dini

social networks and geo-social networks, location based services, database systems or cloud computing. The characteristics and their mapping to challenges can be seen in Table 3.1. In the following sections we provide an analysis of how the existing literature addresses each characteristic, whereas summaries of the approaches found under each challenge are provided per characteristic in Tables 3.2 to 3.5.

3.4 Challenge 1: Profiling, inference and discrimination

3.4.1 CR1 - Prevent inference

The processing of personal data that can reveal more information about a person, such as ethnic origin, political opinions, or even the possibility to uniquely identify a person, is prohibited in GDPR (Article 9). A framework for IoT personal devices uses inference prevention techniques for the protection of user privacy and data, through an Adaptive Inference Discovery Service (AID-S) which provides dedicated functionalities for data control in [286, 288].

These approaches make use of a more general framework which can manage data collected from different devices and whose capabilities can be extended, referred to as a Personal Data Manager (PDM). Further work from the same authors provides a case study on fitness trackers, examining how a third party can obtain and exchange Fitbit data when the user has granted access to that third party [289]. A privacy framework in the domain of fitness, where users can define which of their fitness data cannot be inferred is introduced in [46]. An extension of this model is using a decentralised architecture, where more restrictions and control on possible data combinations can be applied [244]. A quantified self application in IoT is used as a case study in [77], since such systems impose many privacy risks, where privacy risk analysis is performed as the user creates or updates her privacy settings, and inferences are prevented by informing the user about possible risks. IoT privacy assistants that are used in privacy-aware smart buildings systems acquire and enforce user's privacy preferences regarding sensor data, such as occupancy status, location or thermostat readings, and employ privacy specific policy elements to model the user privacy settings in [218]. Another approach is a negotiation mechanism that takes a holistic approach satisfying the requirements both of the user and the third party [8]. The privacy concerns over the potential disclosure of users' identity are addressed in [5] with an interactive tool which models the information shared by users and calculates possible inference risks of data combined by wearables owned by the users and their online social networks accounts. Inferences are prevented by displaying the relevant inferences and risks to the users in the available interface where the user can browse through the various risks in order to get a better understanding. Another approach aiming to allow the users to protect their own privacy in IoT-based systems, follows a number of steps for the protection of privacy in [27]. One of these steps is the assessment of the privacy risks associated with the release of personal data using the Privacy Oracle component, assisting in the prevention of inferences. In the *Privacy-EnhAction* framework we propose and discuss later in Section 3.8, the prevention of inferences is accomplished by a dedicated privacy risk analysis that takes place when a third party makes a data request, and the policy statement of the third party is compared against the users privacy preferences in relation to the data requested.

3.4.2 CR2 - Provide data transformation

Data transformation is a topic that has been discussed a lot since the introduction of GDPR (Recital 26). In the reviewed literature for this characteristic under the first challenge, the

Alexia Dini

authors focus on approaches that are using data transformation techniques to protect users' data privacy against profiling and data inferences risks.

Anonymization techniques are applied to sensitive information before the data are shared in a smart home scenario, where a framework for privacy modelling is proposed for any personal data collected by sensors in [231]. The Inference Discovery component mentioned earlier recommends data transformation when inference probabilities are present in [288]. In the Fitbit case study, the same architecture for data transformation is being used in [289]. Any data protection scheme for data transformation can be used in a smart environment for monitoring patients at home in [26], while appropriate data masking for e-health data satisfying both the user and the third party is achieved in [294]. In [328], the authors address the privacy issue of patients by introducing a storage system for privacy preservation with access control for medical records produced in smart IoT-based healthcare networks. Data transformation techniques are used to encrypt medical files which are then stored in a secure storage system.

A powerful and versatile encryption scheme for the IoT based on attributes is presented in [321], which gives the capability to data owners to manage the credentials of data users in a comprehensive way, addressing data privacy and access control for the protection of data owners. A lightweight secure health storage system for IoT is proposed in [82], which uses data transformation techniques, such as public-key cryptography over symmetrical cryptography, in order to preserve both the privacy and the availability of patient data. A protocol providing full privacy for Location-Based services uses obfuscation techniques for the protection of users identity, location and usage profile in smartphones in [255]. In another approach for location privacy in IoT, an Enhanced Semantic Obfuscation Technique (ESOT) is used for the preservation of user location privacy [296]. The results of the performance evaluation of the technique show that it accomplishes an improved location privacy protection. The privacy in the context of IoT is analysed in [31], where data anonymization is proposed as a mechanism to preserve privacy. While in the PrivacyOracle approach discussed earlier, data release is controlled through the application of data transformation techniques, such as anonymization or data perturbation, before sharing the data in [27]. In [237], the blockchain based solution that assists users by giving them with control over access to their personal data in IoT, the Intelligent Policies Analysis Mechanism provided implements pseudonymisation techniques for the protection of user data. ADvoCATE is presented in [237], a user-centric blockchain based solution that assists users by providing them with control over access to their personal data in IoT. In particular, by the provision of an Intelligent Policies Analysis

Alexia Dini

Table 3.2: Approaches for Profiling, Inference and Discrimination - CH1

Char.	Means	Methods used	Appr.
CR1	PDM	Use of inference probabilities	[286, 288, 289]
	PF	Specification of data categories and purposes for data collection	[46, 244]
	PF	Privacy Risk Analysis, harm trees	[77]
	PF	IoT Privacy Assistants	[218]
	PF	Privacy negotiation mechanism	[8]
	Tool	Visualisation of risks and possible inference available to user	[5]
	PrivacyOracle	Assessment of privacy risks associated with data release.	[27]
CR2	PF	Anonymization techniques	[231]
	PDM	Any data transformation technique	[288, 289]
	PF	Differential privacy, anonymization	[26]
	PF	Perturbation, randomisation, quantization	[294]
	SS	Encryption	[328]
	ACS	Encryption	[321]
	SS	Cryptography	[82]
	LBS	Obfuscation	[255]
	ESOT	Obfuscation	[296]
	General	Data anonymization	[31]
	PrivacyOracle	Anonymization, perturbation	[27]
	ADvoCATE	Pseudoanonymisation [237]	

Abbrev: PDM = Personal Data Manager; PF = Privacy Framework; SS = Storage System; ACS = Access Control System; LBS = Location-Based Services

Mechanism the users can notify third parties for their requests over the deletion of their data. In the *Privacy-EnhAction* framework we propose, any data transformation technique can be applied before data are released.

3.5 Challenge 2: Control and context-sensitive sharing

3.5.1 CR4 - Provide control of personal data to users

The primary aim of GDPR is to give control to users over their personal data and the approaches presented under this characteristic serve this purpose. There are similarities between this characteristic and the “Prevent Inference" characteristic analysed under the first challenge, but the approaches presented here focus on individuals remaining in control of their data.

The risks of the user’s privacy settings regarding fitness and location data are commu-

Alexia Dini

nicated to the user, along with how these risks are influenced by these settings through the available interface in [77]. The Privacy Coach is a mobile phone application that provides users with control on how their RFID smart card data are shared, by comparing the user privacy settings and the third party privacy policies in [42]. Personal Privacy Assistants assisting users to find adjacent IoT systems that may collect any personal data from their IoT devices, are presented in [74,218]. The user can configure her privacy preferences according to the privacy practices of those devices and have control on her data. The *Privacy-EnhAction* framework that we discuss later provides control of personal data to the users by enabling them to specify their privacy preferences for their devices, by comparing the user privacy settings and the third party privacy policies and informing the users about possible privacy risks, providing recommendations that assist the users in making informed consent decisions regarding their personal data.

3.5.2 CR6 - Provide tools for data management to users

The PIM platform in [274] provides users with an interface, where they can perform actions relevant to data deletion or rectification. The architecture described earlier in [26] provides a user interface that notifies the user about possible privacy risks when a new data request is received and suggests suitable actions to reduce them. An interface which includes a privacy preference and a privacy risk pane is provided in [77], also described earlier, aiming to find any possible system privacy risks in the underlying context and informs the user using a simple language, avoiding technical terms. An interface that enables the users to understand and manage the privacy risks in their smart home is proposed in [210]. Through the proposed *Privacy-EnhAction* framework, the users can be notified about potential privacy risks and will be provided with recommendations to reduce them.

3.5.3 CR5 - Provide monitoring and control of devices that collect data

Providing users with control and monitoring of devices that collect data is an important aspect in IoT. The Personal Information Management (PIM) platform gives the users the possibility to take informed decisions regarding their data from their IoT devices in [274]. The users regulate how their data can be used through an interface, by specifying who can access them and why, or by carrying out specific actions on the data. Third parties that collect data can be controlled and monitored using a Personal Data Manager (PDM) in [288], which enables the user to define her privacy preferences to designate how the third party can

Alexia Dini

exploit her data. The IoT Privacy Assistants exploit machine learning methods to create models and specify the privacy settings of the users based on context in [74]. In [218], IoT Resource Registries (IRRs) publish the privacy practices of the devices stored in their database, while the IoT assistants inform users about them, enabling the users to define their privacy preferences accordingly, having control and monitoring of the surrounding devices. In the *Privacy-EnhAction* framework that we discuss later, the users can regulate access to their data by being able to specify their privacy preferences which are enforced by the framework.

3.5.4 CR7 - Provide ability for data erasure

Provision for data erasure is a requirement in GDPR (Article 17), known as *“The right to erasure"*. In the PIM platform in [274], the users can perform selective actions, such as deletion, on the data they produce based on the underlying context. In [74], the Personalised Privacy Assistants inform the users if the surrounding IoT resources provide user configurable settings, such as the ability for data erasure. In [187], PrivySharing, a framework that is based on the blockchain technology for privacy preservation in the smart city is presented which integrates a number of the GDPR requirements, one of which is the deletion of user data after an explicit time. ADvoCATE is presented in [237], a user-centric blockchain based solution that assists users by providing them with control over access to their personal data in IoT. In particular, by the provision of an Intelligent Policies Analysis Mechanism the users can notify third parties for their requests over the deletion of their data. While in [20], the blockchain based privacy preserving framework for healthcare that is presented allows the users to delete their own data at any time. The framework uses an Inter-Planetary File System (IPFS), where the user personal data including health data are separated from public data, and are kept offline.

Alexia Dini

3.5.5 CR8 - Provide transparency

Transparency is required in order for the users to be presented with easy to use systems that make them aware of the privacy implications in a comprehensive way. In [77], the proposed user interface enables the users to inform the third party about their privacy preferences, while being able to visualise the possible privacy risks and the associated impact on their personal data, thus going one step further than CR6. In [315], the authors use six User-Centric-Control-Points (UCCPs) as requirements for the design of privacy-preserving

Table 3.3: Approaches for Control and Context-sensitive sharing - CH2

Char.	Means	Methods used	Appr.
CR4	PF	Risks displayed as the user sets her preferences	[77]
	PC	Comparison between user and third party privacy policies	[42]
	PA	User is informed about each resource's data practices	[74]
	PA	User has full control of personal data	[218]
CR5	PIM	User can visualise and perform selective actions on data	[274]
	PDM	User controls and monitors devices by defining privacy settings	[288]
	PA	User can discover IoT devices and configure her privacy preferences	[74, 218]
CR6	PIM	UI allows users to perform various tasks	[274]
	PF	UI provides privacy risks and recommendations to users	[26]
	PF	UI with privacy preferences pane and privacy risks pane	[77]
	PF	UI enables user to manage privacy risks in smart homes	[210]
CR7	PIM	User can delete her data	[274]
	PA	User can define when data should be erased	[74]
	PrivySharing	User data can be deleted	[187]
	ADvoCATE	Users can request deletion of data	[237]
	PF	User can delete her data	[20]
CR8	PF	User can communicate her privacy settings and visualise risks and impact	[77]
	PF	Balances any potential risks with the possible advantages	[315]
CR9	PDM	Provides optimal privacy settings for balancing user and third party privacy	[288, 289]
	PF	User assesses privacy risks and balances them against possible benefits	[26]
	PF	Negotiation process between user and third party	[294]
	PF	Automatic negotiation process between user and third party	[8]
	ESOT	Distance between user actual location and obfuscated location provides a balance between user privacy and utility service	[296]
	LocationSafe	User has the final say over the release of her location data and the trade-off between privacy and utility	[142]

Abbrev: PF = Privacy Framework; PC = Privacy Coach; PA = IoT Privacy Assistant; PIM = Personal Information Management; PDM = Personal Data Manager;

Alexia Dini

solution in the smart home, that provide transparency by allowing the user to have control over the data collected about them by smart home devices. The six UCCPs also assist users to balance any potential risks with the possible advantages from sharing the data collected.

3.5.6 CR9 - Provide balance of privacy between user and third parties

A requirement in successfully involving the users in the protection of their data is to make them understand the possible risks of data sharing in order to decide whether to take these risks in exchange of potential benefits. The Inference Discovery component proposed in previous works has the capability to recommend optimal settings, which represent how to decrease the inference risk while increasing the amount of shared data [288, 289]. The privacy architecture in [26], described earlier, allows the user to negotiate with the third party, based on the potential benefits she may enjoy, before taking any actual sharing decisions. A negotiation process for providing balance between the privacy of the user and the third party, before reaching an agreement between them, is enforced in [294]. No user interaction is needed in a similar privacy negotiation mechanism in [8]. The ESOT technique, presented earlier [296], provides a balance between user privacy protection and service of utility. The experimental tests showed that the technique provides an acceptable distance between the original location and the obfuscated location, providing a balance between user location privacy and location service utility. The design and implementation of a privacy module for GPS, LocationSafe, is presented in [142], which runs on GPS enabled devices aiming to provide users with granular control over the release of their location. Here, the user is empowered to decide about able to start a negotiation with the third party, where both the user and the third party are provided with a number of recommendations, aiming to provide a balance between their privacy.

3.6 Challenge 3: Consent and uncertainty

3.6.1 CR2 - Provide data transformation

The approaches reviewed in this section aim to realise issues such as user consent and notice, using data transformation techniques. UPESCI is an approach for enforcing privacy in cloud-based services, enabling user consent by using basic data protection functionalities, such as encryption for e-health data [129, 130]. The UPESCI functionality is extended in [214], by specifying the necessary mechanisms for an even communication between the parties involved. The Privacy Manager (PM) used in a smart grid context adopts randomisation based methods and homomorphic encryption schemes for the protection of personal data related to consumers' usage habits [99], while in PASiC, the users can define their consent through the Consent Manager Unit in a cloud-based services scenario, using obfuscation and encryption for data protection [16]. A solution for addressing user privacy concerns is using a "Data Access Manager" (DAM) for controlling access to the data in IoT systems and for the enforcement of the user consent preferences, which are captured by the "Consent Manager" component in [115].

The DAM component controls the actual release of data by performing filtering or mask-

Alexia Dini

ing on the data. Another solution for addressing the problem of user privacy protection in smartphones is the implementation of ProtectMyPrivacy (PmP) for Android, which provides privacy control to the user by distinguishing between data access made by the application developers in their code, and access done by a third party in [61]. The users can allow or deny access to their data based on this information that is presented to them. Private data such as location or contacts are anonymized by the PmP's anonymization module. A European effort, where user privacy in IoT is considered a major quality perspective, empowers the users to be in control of their IoT devices and privacy, through the proposed IoT architecture reference model in [30]. In order to ensure user privacy, sensitive information is encrypted based on dynamic attribute-based policies. Sec4IoT, presented in [69], is a framework for the secure storage of data, aiming to provide enhanced privacy and security to the traditional IoT architecture, which allows the users to get back their privacy rights by specifying their own privacy rules. In the proposed architecture, the data is encrypted with a secret key before being being uploaded to the servers.

3.6.2 CR4 - Provide control of personal data to users

The approaches presented in this section aim to address user consent and notice by enabling the user to be in charge of her personal data. The Personal Data Custodian (PDC) allows the user to specify her privacy preferences and decide whether to accept or reject data requests, while disseminating personal data according to the user preferences in [201]. The Privacy Enforcement Points (PEPs) [129, 130] assist users in protecting their data from unauthorised access, where users can annotate the data making it available only to services that they have authorised. The Privacy Manager [99] covers user consent by allowing users to specify their privacy conditions, making them involved in the personal data privacy management. A privacy framework built on policies for implementing informed consent in a Cooperative Intelligent Transport Systems (C-ITS) scenario and a smart city scenario allows the user to define her privacy settings and preferences, using specific rules, which control how data can be accessed, collected and used [206, 207]. Such data can be vehicle data, like speed or location, or smart space data, like sleep patterns or activity. The ProtectMyPrivacy app presented in [61] provides to the user information regarding who wants to access their personal data through a user interface. The users are then able to decide whether to allow or deny access to their data, keeping control of their personal data.

Alexia Dini

3.6.3 CR6 - Provide tools for data management to users

In this section, the reviewed approaches equip the user with tools enabling informed consent and notice. The Personal Data Custodian [201] gives users the possibility to manage their data through a user interface. The users can employ the provided dashboard for reviewing the declarations of the third parties and give their consent by using the available menus, while they also have the choice to use a simplified adaptation of the privacy language, which has been implemented in a version that uses natural language. The provided user interface enables users to make better informed choices on their privacy settings by specifying their privacy preferences and then the system informs them about the privacy risks of their choices. In the scenarios of C-ITS and smart cities in [206, 207], the authors address the informed consent requirement using an approach based on policies. In [208], the authors propose SecKit, an open source model-based security toolkit that can be used for specifying and enforcing privacy rules. Based on these approaches, the use of tools enable the users to manage properly their data and express their consent.

3.6.4 CR10 - Provide enforcement of user privacy preferences

For a system to be considered as privacy enhancing, it must enable the users to determine how their IoT devices can take actions on their behalf. Enforcing the user privacy preferences before the interaction with IoT services can help enhance the protection of user privacy [130]. The third party sends the request in a policy statement, which is then evaluated against the user’s privacy preferences in order to check if it respects them in [288]. In this scenario, the user privacy preferences are enforced by the system. The specification of user privacy preferences and related enforcement mechanisms are supported to allow the execution of a request only if the privacy policy of the third party complies with the user privacy settings in [46]. Smart objects assist in the enforcement of privacy, by determining the privacy metadata for any new data and by checking if third party privacy policies satisfy the users preferences in [244]. IoT assistants capture user privacy preferences, which are then enforced by smart buildings in [218]. UPESCI, which was described earlier, enables the enforcement of user privacy preferences through Privacy Enforcement Points (PEPs) by the user herself in [129, 130]. The “*Data Access Manager*" (DAM) is used for the enforcement of the user consent preferences, which are captured by the “*Consent Manager*" component in [115]. In the IoT Reference Architecture discussed previously, the system enforces the users’ privacy settings using authorization rules [30]. In the LocationSafe privacy module

Alexia Dini

in [142], the user privacy preferences are enforced by the system regarding access to location data from third parties.

Table 3.4: Approaches for Consent and Uncertainty - CH3

Char.	Means	Methods Used	Appr.
CR2	UPESCI	Encryption	[129,130,214]
	PRM	Randomisation, homomorphic encryption	[99]
	PASiC	Obfuscation, encryption	[16]
	CMS	Data masking	[115]
	PmP	Data anonymisation	[61]
	RA	Data encryption	[30]
	Sec4IoT	Data encryption	[69]
CR4	PDC	User can accept or deny data sharing	[201]
	PF	User takes informed decisions	[77]
	PF	User can annotate data	[129, 130]
	PRM	User specifies data conditions and obligations	[99]
	PF	User specifies access and usage of data rules	[206]
	PF	Usage control policy tailored to context	[207]
	PmP	User can decide whether to allow or deny access to her personal data	[61]
CR6	PDC	UI with drop-down menus for user consent	[201]
	PF	UI informing user of privacy risks	[77]
	PF	Model-based security toolkit for privacy rules specification and enforcement	[206, 207]
CR10	PDM	System enforces privacy preferences	[288]
	PF	System enforces privacy preferences	[46]
	PF	Smart objects perform compliance check and enforce user privacy preferences	[244]

Continued on next page

Alexia Dini

Table 3.4: Approaches for Consent and Uncertainty - CH3 (Continued)

Char.	Means	Methods Used	Appr.
	PA	Smart buildings capture and enforce user privacy preferences	[218]
	UPESCI	Privacy Enforcement Points enforce user privacy preferences	[129, 130]
	CMS	The Data Access Manager enforces user consent preferences	[115]
	RA	Users' privacy settings are enforced through authorisation rules	[30]
	Location-Safe	Users' privacy preferences are enforced by the system	[142]
CR11	PF	Privacy by design, highest level of privacy	[77]
	UPESCI	Privacy by default, default privacy configuration	[129, 130]
	PF	Privacy by default, default privacy profiles	[207]
	PrivySharing	Access control rules	[187]
	PF	Privacy by design, blockchain	[20]
CR12	PC	System reports policy mismatches to the user	[42]
	PDC	Dashboard enables users to take informed decisions	[201]
	PRM	User can take informed decisions	[99]
	PF	User specifies obligations for data usage	[206, 207]
	CMS	User specifies the consent parameters for data usage	[115]
	PrivacyGate	System asks for user consent prior to each transaction	[192]
CR13	PF	Calculation of data sensitivity	[231]
	PDM	Prediction of inference risks	[288, 289]
	PML	Calculation of privacy score for data based on privacy preferences and estimated risks	[116]
	PF	Calculation of privacy risks	[26]

Continued on next page

Alexia Dini

Table 3.4: Approaches for Consent and Uncertainty - CH3 (Continued)

Char.	Means	Methods Used	Appr.
	PF	Privacy risk analysis, identification of privacy harms in a given context	[77]
	Tool	Uses information shared by users to determine how they can be exposed to unwanted leakage of further personal data	[5]
	PrivacyOracle	Semantic Web technologies are used to determine possible data inferences along with their associated privacy risks	[27]
	PF	A Privacy Quantification Framework estimates the privacy risks of sensitive personal data	[281]
CR14	PDM	Dialog based recommendations to users	[288, 289]
	PIM	User is informed for risky data-sharing	[274]
	PC	User is informed about policy comparison result	[42]
	PML	Risks and recommendations presented to user	[116]
	PPRP	Privacy risks presented through the UI	[77]
	Tool	Privacy risks and mitigation techniques are presented to the user	[5]
	PrivacyOracle	Privacy risks presented to the user	[27]
	PF	user-friendly privacy risk indicators are delivered to the user	[281]
CR15	PDM	Creation of user profile by PDM	[288, 289]
	PIM	User sets privacy preferences through the UI	[274]
	PML	Hierarchical questionnaire, privacy classifier	[116]
	PF	User defines the intentions for allowing or denying data sharing	[46]
	PF	User defines the intentions for allowing or denying data sharing	[244]

Continued on next page

Alexia Dini

Table 3.4: Approaches for Consent and Uncertainty - CH3 (Continued)

Char.	Means	Methods Used	Appr.
	PPRP	User sets preferences through a series of questions and alternatives on UI	[77]
	PC	Question and answer wizard on UI	[42]
	PA	User sets privacy preferences through the UI using the developed policy language	[218]
	UPESCI	User sets privacy preferences through default configuration	[129, 130]
	PF	User labels selected data, machine learning algorithms classify data	[151]
	RPM	User sets privacy preferences through UI	[99]
	CMU	User sets privacy preferences through UI	[16]
	CMS	User sets consent preferences through the Consent Manager Component	[115]
	RA	User controls and sets privacy settings	[30]
	Sec4IoT	User sets his own rules for data sharing	[69]
	ESOT	User sets his privacy preferences according to her current position and location	[296]
	Location-Safe	User sets his privacy preferences regarding accuracy and frequency of location data disclosure	[142]
CR16	PF	Policies are captured and sent using a policy language	[218]

Abbrev: PRM = Privacy Manager; PmP = ProtectMyPrivacy; CMS = Consent Management Solution; RA = Reference Architecture; PDC = Pers. Data Custodian; PF = Privacy Framework; PDM = Pers. Data Manager; PA = IoT Privacy Assistant; PC = Privacy Coach; PML = User-centered privacy model; PIM = Personal Information Management; PPRP = Privacy Preferences-Risks Pane; CMU = Consent Man. Unit;

Alexia Dini

3.6.5 CR11 - Provide privacy by design or privacy by default

One of the main requirements of GDPR is to provide “data privacy by design and by default" (Article 25). Ann Cavoukian has designed the original privacy-by-design framework, on which many studies have been based [47]. An example of such a study proposes a framework which can be used for the evaluation of the effectiveness of privacy in current IoT systems in [225]. The user interface provided for the specification of the privacy preferences in [77] is using the privacy by design principle, where the users can skip the selection step when they are setting up their IoT device, giving them the benefit of using the highest level of privacy. The PISCES framework separates the user and the third party needs, where the user becomes accountable for protecting the privacy of her data and the third party turns liable for protecting the data it provides to others [104]. The UPESCI framework provides privacy by default by recommending a default privacy configuration, which can be optionally changed in [129, 130]. In another case presented in [207], the users can select one from the available profiles, which are made up from a number of policy rules. The PrivySharing framework introduced in [187] provides privacy by design through the use of access control rules, that the data owner can use to allow or deny access to their assets. The privacy framework presented in [20], provides privacy by design by storing user data in an encrypted form. and by storing public information on blockchain and private user personal data including health data off-chain. In the *Privacy-EnhAction* framework, the system provides default privacy settings to the users, which the users can keep or change, satisfying the privacy by design GDPR requirement.

Alexia Dini

3.6.6 CR12 - Provide ability to users to make informed consent choices

The Privacy Coach, discussed previously, allows users to take decisions regarding using a particular RFID tag or not, by comparing their privacy preferences to the RFID privacy policy as described in [42]. The Personal Data Custodian (PDC) meets the requirement for informed consent, which is given by the user through her choices in the PDC in [201]. The Privacy Manager agent, also discussed previously, provides privacy feedback to the users, empowering them to provide their informed consent for data sharing [99]. In the implementation of informed consent, the users can define constraints and authorisation rules expressing their decisions for granting or forbidding access to their personal data in [206, 207]. The Consent Manager component which was introduced earlier in [115] is responsible for the collection, storage, and maintenance of user consent. The users are able to define the relevant consent parameters for the applications they wish to use, through a consent template including information such as the purpose of data use. The PrivacyGate proposed in [192] is an extension to the Android operating system which enables the users to control their privacy in a more reserved approach than the existing mobile operating systems. For example, Android requests the user consent before giving access to applications on personal data, but applies the user's decision to later requests. In PrivacyGate, the system prompts the user to give her informed consent before each transaction.

3.6.7 CR13 - Estimate privacy risks of data collection/inference to users

A risk estimation process should consider the user profile, the context and the user's trust in the third party, since privacy risks are closely related to inferences on collected data. The Privacy Risk Detection component described in [231] uses specific functionality to estimate how sensitive data are and the Privacy Management component decides whether the request is allowed or denied. The Adaptive Inference Discovery Service [288, 289] can predict and inform users about the inference risks of a request based on the user privacy preferences. The privacy risks of using a service in a generic IoT environment are estimated in [116], where the risk is considered to be the probability of some actions that can be performed by the service on the user personal data. The Privacy Risk Inference component, that is proposed in [26], assesses the risks of releasing data using factors such as the user profile or the context. A privacy risk analysis methodology described in [77] is taking into consideration privacy parameters aiming to identify and evaluate any possible privacy risks. A tool for gathering the possible inference risks from wearable devices usage was discussed earlier [5], where the information shared by users is used to depict the privacy risks that are present. In Privacy Oracle [27], which was presented earlier, Semantic Web technologies are used in order to determine information about possible data inferences along with the associated privacy risks, such as discrimination or surveillance. An architecture for user-centered privacy risk detection empowers the users to take actions for the protection of their privacy in [281] using advanced machine leaning classifiers and mathematical models. A privacy detection component is used to detect sensitive personal data, while the Privacy Quantification Framework component estimates the privacy risks utilising the sensitive personal data. In *Privacy-EnhAction*, the framework we suggest and discuss later, the system performs a calculation of the privacy inference risks of the requested data in conjunction with the users privacy settings.

3.6.8 CR14 - Communicate risks of data collection/inference to users

Privacy risks of data collection have a potential impact of a privacy breach incident, so it is very important to notify the users about them. The Adaptive Inference Discovery Service can make recommendations to the user about which personal data should not be shared after

Alexia Dini

estimating the risks of data collection [288, 289]. In the PIM platform proposed in [274], the algorithms used include methods for anomaly detection, which are then employed to notify the user in the case of any risky data-sharing with a service. In the Privacy Coach application [42], if the result of the comparison between the user preferences and the RFID policy is a non-match, the user can decide whether to use the tag or not, whereas in [116], the system notifies the users about the estimated risks and their implications, along with the provision of advice regarding the management of these risks. The Privacy Risk Pane in [77] shows the privacy risks of data collection to the user through the available interface for every selection made in the Privacy Preference Pane. The interactive tool developed in [5] presents to the user all the identified risks for the personal data in a list, from which the user can choose a particular risk and the tool will provide further information about that risk and alleviation techniques. In PrivacyOracle [27], the estimated privacy risks are communicated to the users through the user interface, where they are able to take a pragmatic decision about data sharing. Whereas in the architecture for user-centered privacy risk detection presented in [281], the Privacy Risk Communication Manager module is used for the communication of user-friendly privacy risk indicators to the users in order to assist them in the process of decision making. In *Privacy-EnhAction*, the calculated privacy inference risks are communicated to the userd.

3.6.9 CR15 - Provide ability to users to specify their privacy preferences

The specification of the user privacy preferences is essential in order to provide informed consent to IoT devices or services and is addressed through the introduction of user interfaces. To avoid repetitions, the approaches reviewed under this characteristic are summarised in Table 3.4, under CR15. The methods used for the specification of user privacy preferences are either through a user profile, through the provided user interfaces, by question and answer wizards, or through the provision of a default configuration which the users can keep or change.

3.6.10 CR16 - Prevent excessive data collection

An important principle in GDPR is data minimization (Articles 5, 7), which requires that data processing should only use as much data as needed to successfully accomplish a given task. Only one such approach was encountered, although any system that collects and processes

Alexia Dini

Table 3.5: *Approaches for Honesty, Trust and Transparency - CH4*

Char.	Means	Methods used	Approach
CR3	PF	System informs users about privacy risks, privacy settings recommendations	[231]
	PA	Privacy assistants notify users about surrounding resources policies	[74, 75, 218]
	PDC	User interacts with the software tool for data collection awareness	[201]
	PrivacyGate	Users can see which data is transmitted through each transaction with an application	[192]
CR8	PRM	Privacy feedback, monitoring of the usage of personal information	[99]
	PDC	Natural language version of the privacy policy language	[201]
	PIM	Provides transparency for the sharing and usage of personal data	[274]
	UPESCI	User communicates with the system in a human-readable form	[129, 130]
	PmP	Enhanced system transparency for the sharing of personal data	[61]

Abbrev: PF = Privacy Framework; PA = IoT Privacy Assistant; PDC = Personal Data Custodian; PRM = Privacy Manager; PIM = Personal. Information Management;

the minimum required data without indicating it explicitly also satisfies this characteristic. In this approach, excessive data collection is avoided by enforcing the user preferences when collecting data, as seen in [218].

3.7 Challenge 4: Honesty, trust, and transparency

3.7.1 CR3 - Provide user awareness on data collection

The users are informed about privacy risks and are assisted by the recommendation of appropriate privacy settings in the privacy framework presented in [231]. In the smart building scenario in [218], the users are informed by privacy assistants about possible privacy implications in the interaction with the smart building. In the privacy infrastructure in [74, 75], privacy assistants can discover the privacy practices of adjacent devices and make the users aware about their data collection routines for personal data, such as video data. The Personal Data Custodian in [201] interacts with the users making them aware of data collection by third parties. The PrivacyGate system in [192] allows the users to have a clear view of which data is transmitted to a third party through each transaction.

Alexia Dini

3.7.2 CR8 - Provide transparency

Under this characteristic we have reviewed approaches that address transparency by providing user-friendly systems that make the user aware of data collection in a comprehensive form according to the GDPR Article 12(7) [93]. Privacy feedback is also another method that is used to enhance transparency for the user. The Privacy Manager Agent, presented in [99], monitors the usage of personal information and provides privacy feedback to the user about the handling of her personal information ensuring transparency. In the Data Custodian in [201], the user can either consult the declarations of third parties and give her authorisation through the available menus, or use an adaptation of the policy language developed. The approach described in [274] addresses the issue of user trust by providing a transparent communication regarding personal data between the user and the system. In [129, 130], the system provides the user with an interface for data management in the cloud, where she can transparently communicate with the service and modify the service capabilities according to her privacy preferences. Whereas the ProtectMyPrivacy app presented in [61] provides enhanced transparency to the users in order to decide whether to allow or deny access to their personal data to third parties.

3.8 The proposed user-centric privacy protection framework for IoT

Alexia Dini

An important contribution of this doctoral thesis is the suggestion of a generic user-centric IoT privacy protection framework, guided by the user needs for privacy according to the GDPR requirements. The proposed framework, which we call *Privacy-EnhAction*, can be seen in Figure 3.1 and it was created based on the sixteen characteristics extracted from the existing literature and the specific needs for user privacy protection that GDPR calls upon IoT. *Privacy-EnhAction* is built on a number of steps, that are shown in Figure 3.2 and described in more detail in Table 3.6. A more generic approach suggesting a number of points for privacy protection and preservation, as well as conformity to law, was presented in [301] targeting how data should be processed under the GDPR.

Since transparency is a key requirement in GDPR, it must be easy for the users to comprehend how, when and why their data is collected, so that they can provide their informed consent. The provision of a well-designed user interface is considered necessary, in order for the users to manage their personal data. It is suggested that a layered settings interface

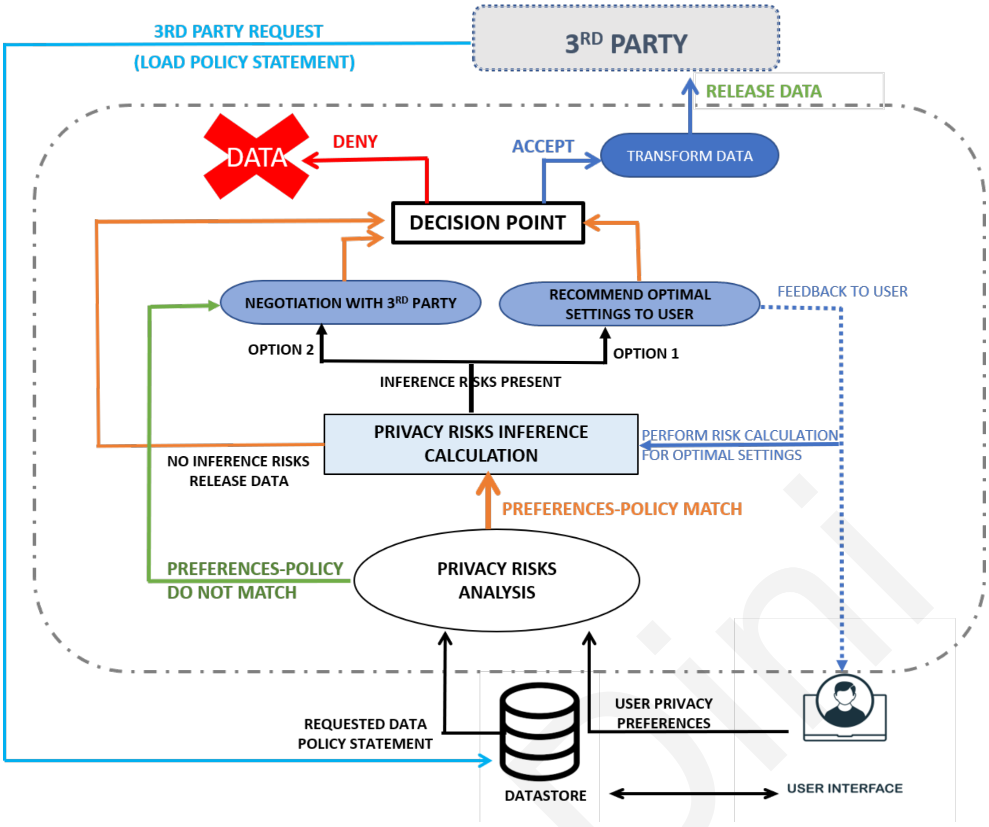


Figure 3.1: Proposed architecture of the proposed user-centric privacy framework, Privacy-EnhAction

Alexia Dini

should be developed as in [19], where the users can take decisions on a less granulated level and have the ability to apply more granular settings when they need more detailed control.

3.8.1 Description of Privacy-EnhAction steps

The steps that constitute the processes of the framework demonstrate how the user can be provided with the functionalities and the tools needed in order to be in control of her personal data created by IoT devices, while they are general enough to be used in any IoT domain. The proposed steps match the elements of the *Privacy-EnhAction* framework.

The **first step** involves the user who can set the privacy preferences for her IoT devices, which are used to regulate access and processing of the data produced. Depending on the setup of the system, default privacy settings could be available which the user could keep or alter - satisfying the privacy by design GDPR requirement - or the user could specify new preferences according to her needs [129, 274]. This step is connected with the fifteenth characteristic, “Provide ability to users to specify their privacy preferences", the eleventh characteristic, “Provide privacy by design or privacy by default", and the tenth characteristic, “Provide enforcement of user privacy preferences".

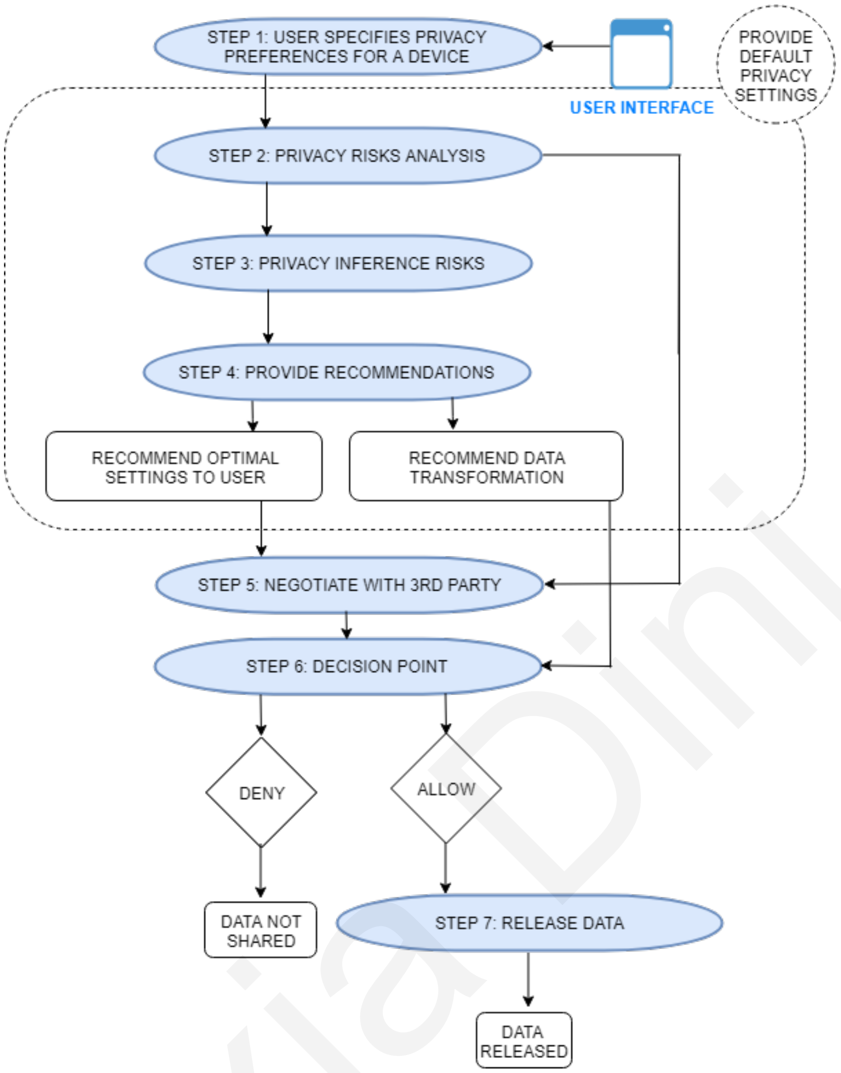

STEP 1: USER SPECIFIES PRIVACY PREFERENCES FOR A DEVICE
USER INTERFACE
PROVIDE DEFAULT PRIVACY SETTINGS
STEP 2: PRIVACY RISKS ANALYSIS
STEP 3: PRIVACY INFERENCE RISKS
STEP 4: PROVIDE RECOMMENDATIONS
RECOMMEND OPTIMAL SETTINGS TO USER
RECOMMEND DATA TRANSFORMATION
STEP 5: NEGOTIATE WITH 3RD PARTY
STEP 6: DECISION POINT
DENY
ALLOW
DATA NOT SHARED
STEP 7: RELEASE DATA
DATA RELEASED

Figure 3.2: Proposed steps for the development of a user-centric privacy framework for IoT

Ideally, the user should specify her privacy preferences for all the devices she owns, and if she desires different settings for a specific device, she should be able to specify them independently. When a third party or service makes a data request, it is important to verify that the third party's privacy policies and data practices satisfy the privacy preferences of the user. In the scope of the *Privacy-EnhAction* framework, it is assumed that the data practices are encoded in the privacy policies of the third party. The privacy policies represent a set of statements that define how the requested data is intended to be used, including information such as the third party id, what data attributes the third party plans to use, the aim of the data collection, the amount of time the data will be held for, or whether the third party will make the user's data available to other services or parties.

To this extent, at the **second step**, assuming that the third party sends the request along with the policy statement, a privacy risk analysis takes place where the policy statement is compared against the user's privacy preferences in relation to the data requested [42], [171].

Alexia Dini

This step is connected with the thirteenth characteristic, “Estimate privacy risks of data collection/inference to users" and the first characteristic, “Prevent inference". If there is an agreement between the two, then the procedure continues with the next step. Otherwise, a dedicated negotiation procedure between the two parties assumes the process [8] (Step 5).

During the **third step**, the calculation of the privacy inference risks of the requested data in conjunction with the user’s privacy settings is performed. This step is also connected with the thirteenth characteristic, “Estimate privacy risks of data collection/inference to users". Many different techniques can be used for the calculation of the inference risks or the probability of them occurring. Several factors must be taken into account for the estimation of risks, such as the user profile, the context, etc. One suggested method is the use of Machine Learning techniques for the prediction of the privacy risks present on the requested data based on historical user data, which are analysed to identify possible inferences about the user, or to use a knowledge base including the requested data, context, risks and policies. Another method is the use of Natural Language Processing techniques in combination with Machine Learning algorithms to process and analyse textual data, such as privacy policies which are provided by IoT manufacturers, in order to extract meaningful insights about the content of those policies regarding data collection processing and sharing practices. If no risks are present, then the data can be released after being transformed to the appropriate format (**Step 7**). However, if risks are detected then the process continues with Step 4.

In the **fourth step**, it is suggested that recommendations are made to the user through the user interface, based on the data requested and the computed risks. This step is connected with the fourteenth characteristic, “Communicate privacy risks of data collection/inference to users" and the third characteristic, “Provide user awareness on data collection". The user will be presented with recommendations for optimal privacy settings aiming to reduce any risks while increasing the amount of data that can be shared, or to proceed with direct negotiations with the third party. The connection with the sixteenth characteristic, “Prevent excessive data collection" is obvious here, as through the optimal privacy settings, minimum data sharing can be achieved. Furthermore, this step is also connected with the eighth characteristic, “Provide transparency", since the proposed framework will enable the users to notify the third parties about their privacy preferences, while being able to be informed about the possible privacy risks. If the user selects to use the optimal privacy settings, she could either go back to Step 3 to analyse the risks again for those, or continue to the **fifth step** for the negotiation with the third party, where both the user and the third party are presented with further recommendations, aiming to provide a balance between their privacy. If one of the

Alexia Dini

Table 3.6: Proposed steps in the Privacy-EnhAction framework

STEP	Prerequisites	Procedure	Possibilities
1: Privacy preferences specification	User interface available. User has logged in the system for the first time, or logs again to alter her settings.	User sets preferences for her devices.	**1.** User keeps default privacy settings. **2.** User alters default settings. **3.** User defines privacy preferences from beginning.
2: Privacy Risks Analysis (PRA)	Third party has made a data request.	**1.** If this is the first time the request is made, the third party has to send its policy statement. Otherwise, it is already available. **2.** Requested data goes through the PRA. **3.** Third party policy statement is compared with user preferences.	**1.** If policies match and are accepted, proceed to **STEP 3**. **2.** If policies do not match, proceed to **STEP 5**.
3: Privacy Inference Risks Calculation	User profile with the privacy preference settings is available.	Calculation of privacy inference risks through various methods (probabilistic model or learning algorithms can be used).	**1.** If no risks present, go to **STEP 6**. **1.** If risks present, proceed to **STEP 4**.
4: Provision of recommendations	Inference risks have been detected for the requested data.	One of the presented options or both: **Option 1**: Recommend optimal settings to user. **Option 2**: Negotiate with third party.	**Option 1** **1.1** Provide optimal privacy settings to user via feedback. **1.2** Go back to **STEP 3**. **Option 2** **2.1** Negotiate with third party. **2.2** Go to **STEP 5**.
5: Negotiation mechanism	Recommendations presented to both entities.	Negotiation with the third party directly with the user, or negotiation with the third party with no user interaction.	Proceed to **STEP 6**.
6: Decision point-Consent Manager		User can review the applications requesting data and the purpose of the request, and may give or deny consent. User may also withdraw consent to an already granted third party request. User will only have to give her explicit consent once for a specific third party, until she manually denies this permission.	**Option 1**: Accept. Proceed to **STEP 7**. **Option 2**: Deny. Data is not released.
7: Perform data transformation	User has accepted to allow data release.	Data is transformed before released.	Data is released.

parties is not content with the recommendation, the negotiation process continues until both parties come to an agreement. This step is connected with the ninth characteristic, “Provide balance of privacy between users and third parties".

The decision point comes in the **sixth step**, where the decision is taken whether to release the data or not. If the data is authorised for release, then the data is transformed and released (Step 7), otherwise no data is released. This step is connected with the second characteristic, “Provide data transformation", as well as the twelfth characteristic, “Provide ability to users to make informed consent choices". Various techniques are suggested in different studies, such as generalisation, perturbation, obfuscation, k-anonymity, etc. [16, 287, 288, 294, 295, 297]. In all steps, the procedure is as seamless for the user as possible, so that the user interferes only when it is necessary.

Regarding the functionality and features of the Privacy-EnhAction framework, the lit-

Alexia Dini

erature review performed provided insights about the relevant existing tools' and technologies' range of capabilities related to data privacy and compliance. However, the proposed framework goes beyond these capabilities by providing a comprehensive set of characteristics specifically tailored to address the challenges of GDPR compliance in the IoT domain. These characteristics encompass user control over personal data, transparency in data collection and processing, purpose limitation, data minimization, security measures, consent management, accountability, and mechanisms for user rights enforcement. The framework integrates these essential components into a cohesive structure that promotes user empowerment and privacy protection. Furthermore, while the reviewed tools offer analysis features to assess privacy risks, the proposed conceptual framework incorporates advanced analysis techniques that focus specifically on GDPR compliance in the context of IoT that is user-centric. It contains methods for assessing data collection practices, identifying privacy risks and vulnerabilities and evaluating the effectiveness of consent mechanisms. Using as a basis state-of-the-art approaches, the framework enables a comprehensive evaluation of privacy practices within IoT systems, providing valuable insights for policymakers, organisations, and individuals.

The proposed conceptual framework builds upon efficiency, scalability, and adaptability in handling IoT data. It addresses the challenges presented by extensive data collection and processing in IoT environments, ensuring that privacy evaluation can be performed in a timely manner. Moreover, even though existing tools provide user-friendly interfaces, the proposed framework aims to ensure a seamless and intuitive user experience, by prioritising simplicity and accessibility in the presentation of privacy-related information to the users. The Privacy-EnhAction framework is also designed to integrate smoothly with various IoT platforms and devices.

Alexia Dini

3.9 Discussion

The characteristics presented in Sections 3.4 to 3.7 aim to serve as guidelines that could act for the benefit of the end users by ensuring that they remain in control of their personal data. They could also assist to thoroughly address principles, such as transparency, for example by providing a user interface through which the user can transparently communicate with the service and modify the service capabilities according to her privacy preferences [129], or data minimization, for example by enforcing the user preferences when collecting data [218], which are considered critical under the directions of the GDPR.

Table 3.7: Challenges and techniques used in literature related to the characteristics

	Challenges					Techniques Used							
Char.	**CH1**	**CH2**	**CH3**	**CH4**	**Total**	**PL**	**ML**	**RM**	**DA**	**NP**	**BC**	**Other**	**Total**
CR1	8				8	2	7		1	1		2	13
CR2	6		5		11	2	2			2	1	3	10
CR3				3	3	2	1	1				1	5
CR4		4	7		11	5	1					1	7
CR5		4			4	1	3						4
CR6		4	4		8	3	1	1		1	1		7
CR7		5			5		2				3		5
CR8		2		5	7	3	2				1	1	7
CR9		5			5		2					3	5
CR10			6		6	3	1		1			1	6
CR11			6		6	3	1				2		4
CR12			5		5	3						2	5
CR13			6		6		3	1		1		1	6
CR14			6		6		4					2	6
CR15			14		14	3	5		1			3	12
CR16			1		1	1							1
Total	14	24	61	8	106	31	35	3	3	5	8	20	103

PL = Policy Language, ML = Machine Learning, RM = Risk Modelling, DA = Data Analytics, NP = Negotiation Protocol

Alexia Dini

Through the characteristics, the provision of dedicated functionalities, such as for the specification of user privacy preferences or the restriction on which portions of personal data cannot be inferred, become possible for the protection against data inferences. Additionally, the primary aim of GDPR to give control to the user over her data and privacy is taken care of through the estimation and communication of the privacy risks to the user, allowing her to decide whether to allow or deny access to her data, achieving the requirement for informed consent, or by the recommendation for providing tools, such as user interfaces to the users. The GDPR requirement for the provision of data Privacy by Design or by Default is also met through the characteristics, where it is suggested that through available user interfaces, default privacy settings become accessible to the user, that can be changed according to the needs.

Practitioners or developers can embody some or all of the characteristics in the design of privacy frameworks to contemplate user privacy for IoT applications avoiding the exposure of personal data without clear and explicit permissions, whilst the users will be empowered with control over their data and privacy using the information provided by the framework, such as data sharing inference risks [288], to their benefit.

All the approaches reviewed in this chapter are summarised in Table 3.7. In this table, we can see the number of times each characteristic has been addressed in the existing literature under each challenge.

For example, CR1 has been addressed in 8 of the reviewed approaches under Challenge 1, while CR2 has been addressed in 6 of the reviewed approaches under Challenge 1 and in 5 approaches under Challenge 3. The techniques employed to satisfy each characteristic can also be found, along with the number of times each technique is encountered. For example, Policy Languages have been encountered in 2 approaches that address CR1 and Machine Learning techniques have been used in 7 approaches that address CR1. The total numbers displayed do not match the number of reviewed papers (45), since each paper may address more than one characteristic or challenge. From this summary, the following important findings can be extracted:

1. From the comparison of the total number of approaches tackling each challenge and characteristic, we observe that the third challenge, “*Consent and Uncertainty*", outperforms the others in terms of received research interest in the area of user privacy in IoT. This is encouraging, as informed consent is key in enabling users to disclose their data without compromising their privacy, while it shows that the need for more advanced models of informed consent has been recognised by researchers, addressing the distinct characteristics of the IoT. The solutions proposed throughout the literature aim to make user consent more efficient without the use of extra needless constrains. Under the GDPR, informed consent calls for the awareness and perception of the user of how her data are collected and exploited, along with the need for the user understanding the pros and cons related with the use of her data. Due to the pervasive features of IoT, this is becoming the greatest challenge, since data may be collected without the user awareness.
2. The second challenge, “*Control and context-sensitive sharing of identity*", has also received the researchers’ interest in the reviewed literature, echoing the GDPR requirement for handing control to the users over their personal data and privacy. The presented solutions aim to empower the users with the provision of the necessary tools

Alexia Dini

to control and manage their data, such as the definition of constraints and policies.

3. The fourth challenge, “*Honesty, trust, and transparency*", has received the least research interest compared to the others. This is quite expected, since GDPR has been recently introduced and such notions were not primarily considered in privacy preserving approaches so far. Transparency aspects have been addressed in other areas, such as recommender systems [22], however, even though interdisciplinary analysis can draw a picture on many transparency aspects to empower users with more control over their personal data in the IoT domain, such studies constitute only a minority [243].
4. The most addressed characteristic is CR15, “*Provide ability to users to specify their privacy preferences*", proving that the reflection of the users’ preferences in a system enhances privacy. As people often do not know their privacy preferences in a given context, specific privacy harms may occur that are not always clear to the typical user [1]. For this reason, researchers have put a lot of effort in designing solutions which enable the users to understand the context in which their data are collected and are able to specify the suitable privacy preferences. In the reviewed literature it is evident that the availability of the specification of privacy settings to the user becomes more common mainly due to the requirements of regulations, such as the GDPR [75].
5. The second most addressed characteristics are CR2, “*Provide data transformation*" and CR4, *“Provide control of personal data to users".* For CR2, this finding shows that the GDPR recommendations for such data security measures have been taken seriously by the research community. Furthermore, it is obvious from the techniques used for data transformation in the reviewed literature, that data anonymisation has been preferred, as the results of this process are permanent and cannot be reversed, addressing the concern that if data are not anonymised then it is likely that user tracking, data linking or prediction of user future actions can occur [118]. Regarding characteristic CR4, users are empowered with control of their personal data by being able to set their privacy preferences, by providing them with information about who wants to access their personal data, and by enabling them to decide whether to allow or deny access to their data, keeping control of their personal data. This way the users can exercise their privacy rights, while service providers will be able to handle users’ personal data in an efficient and trustworthy way.

In relation to the techniques being used in the reviewed literature, the following observations can be made:

1. Machine Learning techniques have been mostly used in the approaches for the com-

Alexia Dini

putation of risks probabilities, the determination of the purpose and the granularity of data sharing, or for the prediction of user privacy preferences and the automatic configuration of settings. Seven privacy preserving approaches that use Machine Learning techniques have been reviewed. Solutions that are based on Machine Learning can enhance user privacy without revealing personally identifiable information that may lead to inferences, while being able to perform analysis on enormous data sets without the need to examine the data portion of each packet, where sensitive personal data are stored. Therefore, Machine Learning techniques can be considered as a feasible approach to the problem of user privacy and data protection in IoT.

2. The use of policy languages for the specification of user preferences and for expressing complicated policies and rules has also played an important role in the reviewed approaches. These languages aim to model GDPR-compliant privacy policies in order to enable processing of data in such a way that these privacy policies are enforced for the preservation of user and data privacy. Furthermore, a best practice shown in some of the reviewed approaches is that such policy languages are enriched with formal semantics facilitating the creation of better user interfaces for the interaction with the user.
3. Blockchain techniques also appear in the list of the techniques used for privacy protection in the literature reviewed. Blockchain brings many advantages when it comes to security, such as data encryption, decentralisation, or authentication, which can assist in the ongoing efforts for solving the privacy issues of IoT. The distributed structure of IoT makes possible the use of blockchain technology for creating decentralised applications with privacy-preserving transactions across the involved parties. For example, the challenge of sharing data across heterogeneous IoT devices in a secure way can be addressed by the blockchain-provided data immutability feature, bringing high data integrity. The transparency of transactions related to user data improves the user trust in the system, while anonymity enables the user to hide her identity and personal data [94]. Additionally, the use of smart contracts can enhance privacy as these are designed to execute encoded rules echoing the user privacy preferences or privacy-preserving policies, specifying the conditions that have to be met by the third party for the handling of user personal data.
4. Techniques, such as Risk Modelling and Negotiation Protocols, are exploited in a smaller number of approaches to evaluate the privacy risks of data sharing in the IoT context, to give more control to the user and to provide solutions acceptable and advan-

Alexia Dini

tageous to both the user and the service. The assessment of privacy risks and possible implications is used to inform the users about the present risks in their interaction with IoT devices and assist in the specification of user privacy preferences. Due to the diverse areas of applications in IoT and the heterogeneity of devices and produced data, negotiation mechanisms are proposed as the solution for providing a utility-privacy trade-off in IoT data management.

5. Other techniques used in the reviewed literature employ specific data infrastructures for networking monitoring or for securing the communications between the entities involved. Privacy Enhancing Technologies (PETs) are also adopted using tools and mechanisms for providing data anonymization, limitation in the collection of personal data and providing more control to the user.

In the rest of this thesis, our research focuses on two of the characteristics we defined in Section 3.3, namely Characteristic 13, *“Estimate privacy risks of data collection/inference to users"*, and Characteristic 14, *“Communicate risks of data collection/inference to users"*. The decision to focus specifically on characteristics 13 and 14 was driven by the identification of a research gap in the existing literature and the significance of these characteristics in addressing the privacy challenges in IoT, particularly in relation to user awareness.

While in the literature researchers have been actively involved with providing privacy-preserving solutions to address the privacy challenges of IoT, limited attention has been given to the development of user awareness mechanisms that can assist the users in understanding how the data created by their smart devices can be exploited for the extraction of inferences regarding their daily activities and lifestyle in general. This research gap presented an opportunity to explore deeper into these two specific characteristics and contribute to the understanding and development of effective strategies and mechanisms for promoting user awareness of privacy risks.

As the primary aim of GDPR is to give control to the users over their data and privacy, the objective of the work presented in the next chapters is to address this requirement, by empowering the users with control over their data, making them aware of the possible inference risks from data collection and sharing by their smart home devices and fitness trackers and therefore providing them the chance to decide whether to allow or deny access to their data. Therefore, we define the second Research Question that this thesis aims to answer as follows: **RQ2:** ***“What inferences can be made from data collected from smart home devices and fitness trackers?"***. Chapters 4 and 5 are dedicated towards this aim.

Alexia Dini

Chapter 4

Inference detection in a smart home scenario

Alexia Dini

In the smart home, vast amounts of data are being collected via various interconnected devices. Although this assists in improving the quality of life at home, often the users are not aware of the details concerning data collection apart from the information available on the provider privacy policy. Inference detection is an important area of research in the field of smart home technology. With the increasing use of Internet of Things devices in the home, it has become possible for these devices to collect and analyse large amounts of data about users' behaviour and habits, and as such it is important to put the users inside this loop of information, so that they are well informed on possible uses of the data and the potential risks that this may entail. This Chapter explores the concept of inference detection in a smart home scenario, in our effort to investigate whether the exploitation of data generated by smart home devices can lead to inferences about the occupants routines or other sensitive information. Using machine learning techniques, we aim to draw conclusions about the user routines or activities, providing an answer to Research Question **RQ2**: ***"What inferences can be made from data collected from smart home devices and fitness trackers?"***, in order to inform the users about the findings concerning data inferences through a dedicated web application.

In this Chapter we introduce a data inference framework using the data generated by a smart water meter and motion sensors deployed in a house, and by employing a number of machine learning techniques we aim to test whether such inferences are indeed possible. The results of the process are utilised in the PrivacyEnhAction privacy tool that we briefly present, which aims to inform the user about possible privacy vulnerabilities stemming from her smart home devices and fitness trackers data. The main contributions of this Chapter are: (i) A data inference testing framework tailored to the smart home, an approach which is able to inform the users about potential unwanted inferences, allowing them to perform appropriate adjustments in order to prevent them, (ii) a proof of concept web application that implements the presented approach, along with a preliminary user evaluation, and (iii) the data created and used in the experiments, that are publicly available in the Zenodo portal[1] for replication purposes. The work presented in this chapter is based on our research published in the Proceedings of the 2021 IEEE SmartWorld Congress [161].

4.1 The Smart Home scenario

In this section we describe our experimental setup for the smart home scenario. We have deployed a series of smart devices in a two-floor house where a family of four lives. We installed three multi-sensor Arduino-based nodes that collect information about temperature, humidity, light intensity, sound level and motion at 10 second intervals, and a single smart water meter device that collects real time water flow and consumption information at 30 minutes intervals. Two Arduino-based nodes were installed on the ground floor capturing the two house entrances as well as the kitchen and living room areas, while the third node was placed on the upper floor master bedroom, capturing also the bathroom entrance. Hereafter we will refer to these nodes as “*motion sensors*". The data collected by the smart water meter consists of intelligently pre-processed low-resolution images to enable digit recognition of the water meter reading, utilizing sophisticated low-level feature extraction algorithms. A 3D image of the smart water meter used in the experiment can be seen in Figure 4.1.

Alexia Dini

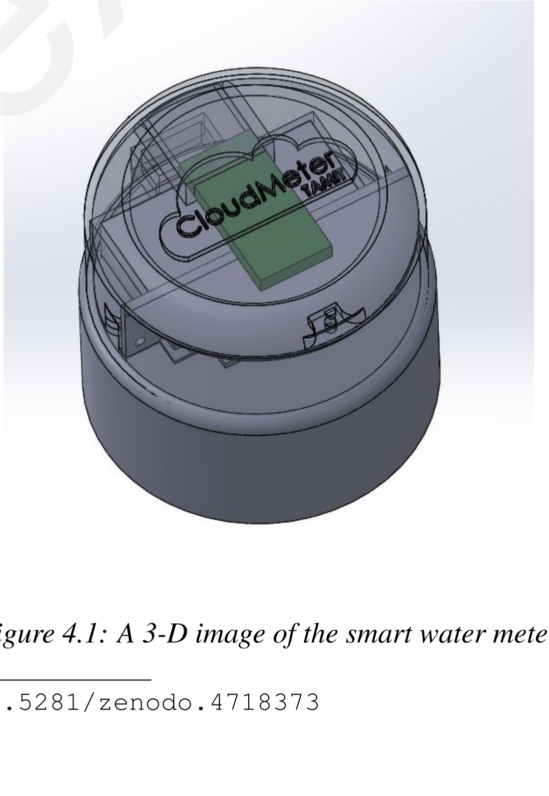


Figure 4.1: A 3-D image of the smart water meter

[1] http://doi.org/10.5281/zenodo.4718373

In order to assist the inference detection process, we have reviewed the state-of-the-art literature in order to derive a list of possible inferences that form a threat to user privacy in our smart home scenario, which can be seen in Figure 4.2. The highlighted parts in the figure designate the inferences we focus on in this part of the thesis. We have left out the most obvious privacy threat, i.e. occupancy detection, as this is an area that has received a lot of research interest [54, 153, 154].

Under this scenario, we aim to answer the following questions: (i) Can we infer the time the residents wake up? (ii) Can we infer the usual time they go to sleep at night? (iii) Can we get insights as to whether they wake-up during the night? (iv) Can we tell which time they leave for work in the morning? (v) Or the time they return?

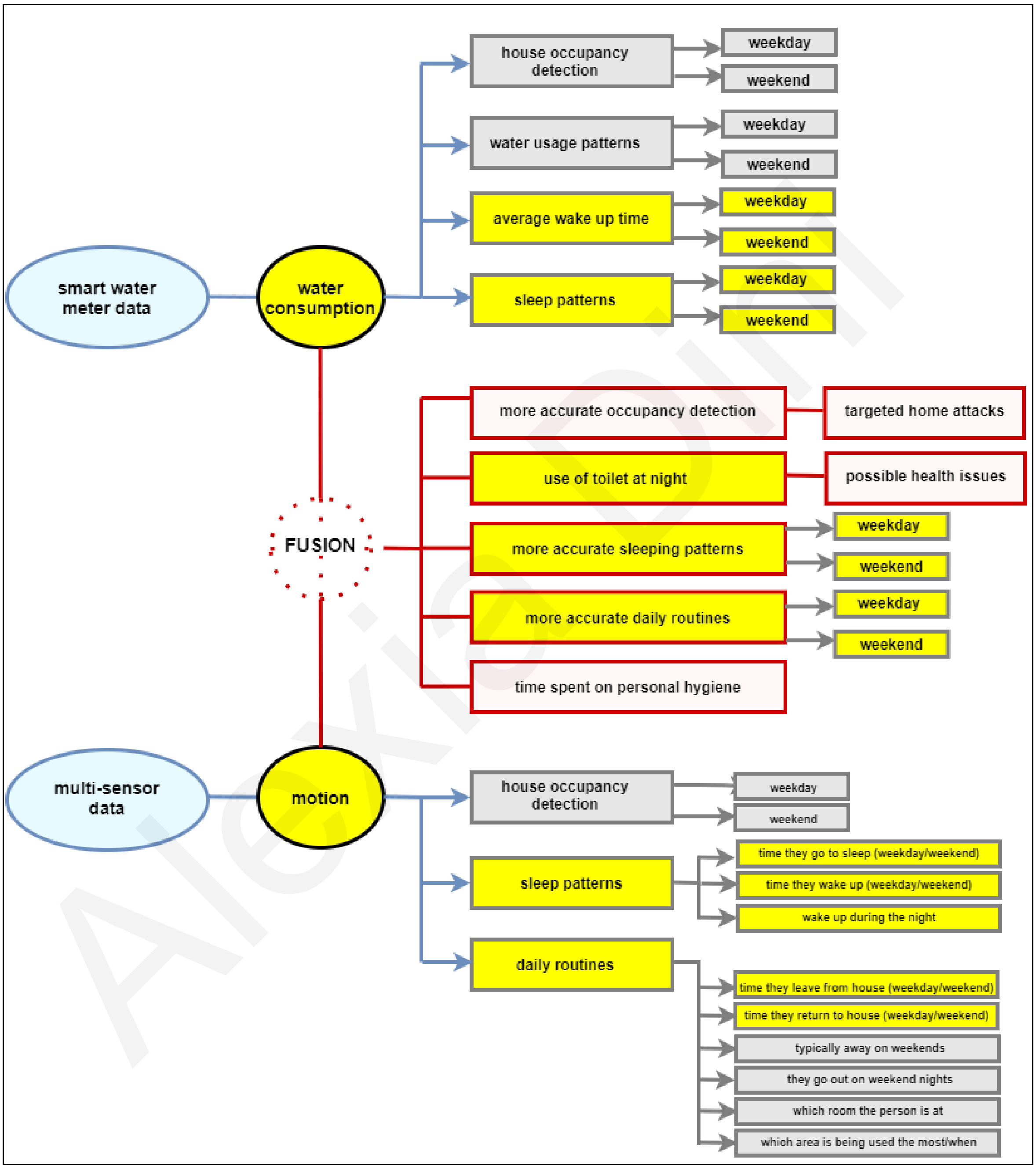
4.2 Methodology

In this section we describe the methodology we used in order to examine and analyse the data in the smart home scenario, using data from the smart water meter and the three Arduino-based motion sensors. Even though we apply this methodology on a specific setting, the described approach can be replicated in other scenarios with similar sensor availability, for example fitness trackers that collect personal user data.

4.2.1 Overview

The methodology consists of four steps. In the first step, we examined and cleaned the sensor data for analysis. We have identified two approaches for organising and analysing the data, which are described in Section 4.2.2. In the second step, clustering analysis was performed on the data, while in the third step the clustering results were exploited focusing on their interpretation, in order to give a meaning to the clustered data and identify what inferences can be drawn. For this reason, a decision tree model was trained to generate a set of rules, which can be used in combination with the insights obtained from visual observations of the data. In the last step, the clusters defined in Step 2 were used as new features in the training data in order to predict to which cluster new data points are assigned to. Based on this, new data was processed aiming to detect if a number of inferences can be drawn from them. Then the users can be informed about the identified inferences, in order to take action, as for example by changing their privacy preferences or settings on their device, as these affect the data collection [84]. The methodology was applied in the following cases/experiments: (1)

Alexia Dini



smart water meter data
water consumption
house occupancy detection
weekday
weekend
water usage patterns
weekday
weekend
average wake up time
weekday
weekend
sleep patterns
weekday
weekend
FUSION
more accurate occupancy detection
targeted home attacks
use of toilet at night
possible health issues
more accurate sleeping patterns
weekday
weekend
more accurate daily routines
weekday
weekend
time spent on personal hygiene
multi-sensor data
motion
house occupancy detection
weekday
weekend
sleep patterns
time they go to sleep (weekday/weekend)
time they wake up (weekday/weekend)
wake up during the night
daily routines
time they leave from house (weekday/weekend)
time they return to house (weekday/weekend)
typically away on weekends
they go out on weekend nights
which room the person is at
which area is being used the most/when

Figure 4.2: Possible inferences from smart water meter and motion sensor data

Experiment 1: Detect inferences using information from motion sensors; (2) Experiment 2: Detect inferences using information from smart water sensors; and (3) Experiment 3: Detect inferences by combining information from motion and smart water sensors.

4.2.2 Data processing and cleaning

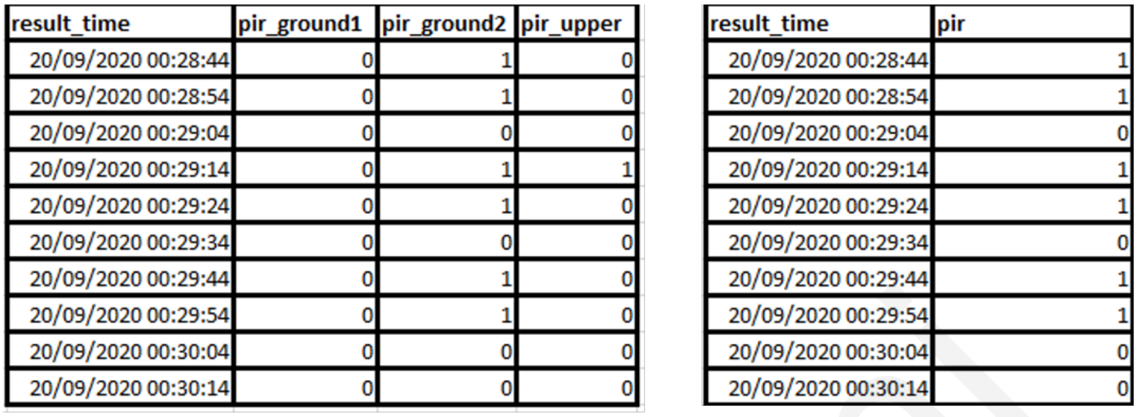
The data collected from the Arduino-based sensors were exploited using the PIR (Passive Infrared Sensor) or motion value created at 10-second intervals, which is 1 when motion is detected and 0 otherwise. It was assumed that motion is detected when we have two consecutive positive PIR values. This means that since data are collected every 10 seconds, motion is indicated when we have two consecutive values of 1, i.e. a period of 20 seconds.

At the beginning of the data collection period and for the duration of two weeks, we asked the household residents to manually record motion data when passing by the nodes (whose detection area falls inside a 110-degree cone with a range of 3 to 7 meters). By examining the sensor data combined with the ground truth values for the same time periods, we came to the conclusion that any single positive PIR value recorded within the period of two minutes is considered as noise and has to be cleaned. In our smart home scenario, the placement of the three motion sensors was strategically chosen to ensure extensive coverage of key areas within the house, including the entrances, kitchen, living room, and upper floor master bedroom. These areas are typically high-traffic zones where human activity is frequent. Given the strategic placement and the sensitivity of the motion sensors, it is expected that motion events will be detected multiple times within a short duration, such as within a two-minute timeframe. This is due to the nature of human movement, where individuals may pass by the sensors multiple times or engage in activities that trigger motion detection. Considering the practicality and real-life usage of the smart home environment, it is important to account for such scenarios and avoid false positives. Therefore, in our data collection process, we set a threshold to filter out and remove any single positive motion detection value recorded within a two-minute period. By doing so, we ensured that only meaningful and significant motion events were captured and considered in our analysis, eliminating noise or insignificant motion detection that may occur.

The datasets from the three Arduino sensors for the same period of time had to be merged, and the PIR values of the three datasets had to be incorporated into a single PIR value (if at least one positive PIR value is present, then the value is positive), in order to identify if motion takes place at the monitored areas during that period. An example taken from

Alexia Dini

the motion sensors dataset can be seen in Figure 4.3. For the smart water meter dataset the processing required was to delete any records containing missing or null values, and to remove any outliers that were identified.



result_time	pir_ground1	pir_ground2	pir_upper
20/09/2020 00:28:44	0	1	0
20/09/2020 00:28:54	0	1	0
20/09/2020 00:29:04	0	0	0
20/09/2020 00:29:14	0	1	1
20/09/2020 00:29:24	0	1	0
20/09/2020 00:29:34	0	0	0
20/09/2020 00:29:44	0	1	0
20/09/2020 00:29:54	0	1	0
20/09/2020 00:30:04	0	0	0
20/09/2020 00:30:14	0	0	0

Dataset example after merge

result_time	pir
20/09/2020 00:28:44	1
20/09/2020 00:28:54	1
20/09/2020 00:29:04	0
20/09/2020 00:29:14	1
20/09/2020 00:29:24	1
20/09/2020 00:29:34	0
20/09/2020 00:29:44	1
20/09/2020 00:29:54	1
20/09/2020 00:30:04	0
20/09/2020 00:30:14	0

Dataset example after merge and pir calculation

Figure 4.3: Example from the motion sensors dataset

As mentioned earlier, two approaches have been identified for analysing our data. The first is the NAZ approach, which identifies the periods when no activity or water consumption takes place, depending on which experiment is being run. The second approach takes advantage of the PIR or motion values from the motion sensor datasets or the consumption values from the smart water datasets, in order to identify the duration that an event (motion, water consumption) lasted for. In the next paragraphs the two approaches are described.

NAZ approach:. In this approach we aim to identify the periods when there is potentially no activity in the house (i.e. the residents are away or sleeping), which we call "No Activity Zones" (NAZ). Based on the data observations and the ground truth data, it is assumed that, for the first experiment using the Arduino nodes, a NAZ exists when no motion is detected for more than 30 minutes (or 180 continuous PIR zero values), i.e. any activity lasting less than 30 minutes is ignored. This is valid when the condition is satisfied for the same time period for the three Arduino-based nodes installed in the house. A number of calculations is performed for summing up continuous values of zero, used to identify the time periods where the sum of continuous zero values is bigger than or equal to 180, indicating a NAZ. Based on this, the start and end time of each NAZ can be computed. In the second experiment with the smart water meter data, it is assumed that when no water consumption is read in the last two readings, i.e. a period of one hour, then in this condition a NAZ is detected. Afterwards, a number of calculations is performed in order to find the start and end time of each water

Alexia Dini

meter data NAZ. In the third experiment where we combine the datasets from the Arduino nodes and the smart water meter, first the NAZ for the Arduino nodes data is calculated, and then a procedure is run to find the corresponding time periods for the smart water meter data against the calculated NAZ periods. As the smart water meter collects data every 30 minutes, the start and end time periods in the water dataset are computed that coincide with the NAZ periods identified earlier by carrying out numerous calculations. The total water consumption that is accounted for during that time is also calculated. The pseudo-code for this process can be seen at Algorithm 6.

Activity-based or consumption-based approach:. In the second approach the values of the motion sensor embedded in the Arduino nodes are used for the first experiment. For the second experiment, the consumption values of the smart water meter are employed, while in the third experiment we perform our analysis using a combination of both datasets. In all cases, hourly resampling on the data is performed in order to combine the different sampling rates of the datasets into one hour intervals. Afterwards, the sum of values for each hour is calculated assembling a daily profile of activity, water consumption, or both.

Alexia Dini

4.2.3 Cluster analysis

In this step, cluster analysis was performed on the three datasets: motion sensor data, smart water meter data, and the combined dataset of motion sensor data and smart water meter data. The objective was to group the data points in a meaningful way based on their similarities and extract valuable knowledge about the family living in the specific household. Since the datasets used in this analysis were not labelled, unsupervised machine learning methods have been adopted, which constitute a prevalent approach for extracting valuable information from unlabelled raw data [101]. By applying clustering algorithms, we aimed to identify patterns and relationships within the data without the need for predefined labels or target variables.

First, the cluster analysis was conducted on the motion sensor data. Various features such as time of activity, duration, intensity, or frequency of motion were taken into account during the clustering process. By grouping similar motion data points together, we aimed to identify distinct behavioural patterns or activity profiles within the household. This analysis provided insights into the family's daily routines, occupancy patterns, and lifestyle preferences.

Next, the smart water meter data underwent cluster analysis. Features such as water consumption patterns, usage frequency, and volume of water consumed were considered during

Algorithm 6 NAZ Algorithm for combined dataset

```
Data: Motion sensor data, smart water meter data Data: $m, w$
Result: Data containing NAZ periods and total water consumption for each period
$M \leftarrow m, W \leftarrow w$
Read motion sensor data $M$
while $M \neq empty$ do
    for each M[$i$]  Check pir value at M[$i + 6$] and M[$i - 6$]
    if $pir\ at\ M[i + 6] = 1\ OR\ pir\ at\ M[i - 6] = 1$ then
        set pir value at M[$i$] =0
    else
        keep pir value
    end
    Group data with cumulative unique count $c$ on zero pir values
    Flag $M[i]$ where $c$ ≥180
    Find NAZ start time $M[i].nst$ and end time $M[i].net$ for flagged data

    Read smart water data $W$
    while $W \neq empty$ do
        for $row\ in\ M[i]$ do
            Get computed $M[i].nst$ and $M[i].net$, $startNAZ = M[i].nst$
            $endNAZ = M[i].net$
            set $counter = 0$

            for $row\ in\ W[i]$ do
                if  $index\ of\ W[i]\ != (length(W)-1)$ then
                    $a$ = reading time of current record W[$i$].$readingTime$
                    $b$ = reading time of next record W[$i + 1$].$readingTime$

                else
                    $a$ = W[$i$].$readingTime$
                    $b$ = W[$i$].$readingTime$

                end
                if $(startNAZ > a\ and\ startNAZ < b)$ then
                    water start NAZ = b
                    StartNAZConsumption = W[$i + 1$].$consumption$
                else if $startNAZ < a\ and\ row\ index=0$ then
                    water start NAZ = a
                    StartNAZConsumption = W[$i$].$consumption$
                if $(endNAZ > a\ and\ endNAZ < b)$ then
                    water end NAZ = a
                    EndNAZConsumption = W[$i$].$consumption$
                else if $endNAZ < a\ and\ row\ index=0$ then
                    water end NAZ = a
                    EndNAZConsumption = W[$i$].$consumption$
                counter = counter + 1

            end
        end
        TotalConsumption for period = EndNAZConsumption - StartNAZConsumption

    end
end
```

Alexia Dini

the clustering process. By grouping similar water usage data points together, we aimed to identify distinct patterns of water consumption within the household. This analysis provided insights into the family's water usage habits or any anomalies in water consumption.

Finally, the cluster analysis was performed on the combined dataset, which incorporated both motion sensor data and smart water meter data. By considering the features from both datasets, we aimed to uncover any interdependent insights or correlations between the family's activities and their water consumption patterns. This analysis enabled a comprehensive understanding of the relationship between motion patterns and water usage within the household, offering valuable knowledge about the family's behaviours, routines, and resource consumption patterns. By analysing the resulting clusters from each dataset individually, as well as from the combined dataset, we obtained a comprehensive view of the family's activities, water usage patterns, and their relationships.

Three unsupervised algorithms were used and their results compared to decide which one fits better in the smart home scenario. First, the K-Means algorithm was applied, one of the most popular unsupervised algorithms [122], in order to find daily patterns based on motion count, consumption count, or both, for the specific household. The second algorithm evaluated was the DBSCAN algorithm, which is a frequently used clustering algorithm [251], that works based on the density of the data points. The third algorithm examined was the Agglomerative Hierarchical Clustering algorithm, which uses a bottom-up approach in grouping the data points, based on similarity metrics. The results of the three algorithms were then compared in order to determine which one is more suitable for the extraction of inferences from the data that could pose a potential threat to users privacy.

4.2.4 Interpretation of Clusters

After our data were clustered, the next step was to interpret the clusters, by analysing and figuring out the data patterns. Visualisations assisted in getting an understanding of the clusters' distribution, by plotting the data using colour. We reinforced our effort by creating and training a decision tree model that uses the dataset features and the clustering result as the label, and produces leaves that provide information about the cluster data. Then the results of the decision tree were combined with visual observations of the data clusters to generate a list of rules for all the scenarios.

Alexia Dini

4.2.5 Inferences detection

The clusters derived in the second step of the methodology were used in combination with the rules generated in the third step for the identification of possible inferences that can be obtained from the data.

4.3 The Smart Home Experiments

4.3.1 Data set description

The Arduino-based nodes collect data every 10 seconds and the smart water meter every 30 minutes. The time-stamped data contains the readings of the sensors embedded in each node, i.e. time or reading, temperature, humidity, brightness, distance from sensor, sound level and passive infrared sensor value, while each record of the smart water meter data contains the cumulative value of water consumption, and the time of reading.

4.3.2 Exploratory analysis

In order to have a more precise model, we ran three cycles of data collection during a period of 6 months, from October 2020 to March 2021. Through this process we have collected approximately 606,000 smart home motion sensor records and 3,100 smart water meter records. The datasets gathered unconstrained real-life context echoing the challenges involved in such an environment, and they were measured with the users operating without any instructions from the research team, as they would perform their daily activities normally.

The first cycle of data collection was the first two months, where data was collected 24 hours a day. The analysis of these results helped us to understand the data. We observed that the data collected was sufficient for the determination of possible privacy inference risks, however we acknowledged the actuality that the results would be based on and limited to the routines of the specific household. Thus, for more comprehensive results in a real world setting, the users would have to be further involved in the process by providing feedback to the system regarding their daily routines. Furthermore, in order to get meaningful results from the data, we had to study the available data so as to get a deeper understanding of how the sampling rate of the data may affect the extraction of inferences and figure out which data points had to be removed, as explained in Section 4.2.2.

The second data collection cycle was between months 3 and 4 and we used the data

Alexia Dini

collected for training and testing purposes. We carried out comparative experiments using the three clustering machine learning algorithms mentioned earlier, namely K-Means, DBSCAN and Hierarchical Agglomerative Clustering. The K-Means algorithm requires in advance the value of K, i.e. the number of clusters. In our case, we utilised K-Means++ that provides better initial seeding for finding the best clusters. DBSCAN does not require the number of clusters beforehand, but uses two other parameters that have to be tuned (epsilon value and minimum number of samples). Agglomerative Hierarchical clustering differs from the other two algorithms in that it begins with each data point being an individual cluster and ultimately creates larger and larger clusters until only one cluster is attained.

Finally, in the third and last cycle, we used the data collected during months 5 and 6 for the validation of our models and the extraction of inferences based on the clusters created and data processing. Following this procedure, all data points were assigned to the the cluster it has been predicted to belong to. As each cluster represents a set of rules that the underlying data conform to and has been defined for each scenario, we proceeded with the processing of the data in order to provide answers to the questions we set earlier.

Alexia Dini

4.4 Experimental results

All the algorithms were implemented using the Scikit-learn library in Python 3.9.0, while the experiments were carried out on an i5 2.5 GHz machine with 4GB RAM.

4.4.1 Comparison of clustering algorithms

The data collected during the second phase of the experiment were examined and used to create machine learning models that can identify patterns in smart water meter and motion sensor time series data and assist in the extraction of inferences. The models were created and tested utilising the K-Means, DBSCAN and Hierarchical Agglomerative Clustering algorithms. In order to evaluate which algorithm works best for our data, we used the three algorithms to find patterns in the daily profiles of the specific household for each scenario, approach and dataset combination. Our intention was to verify if and which of the identified inferences (Figure 4.2) can be determined through the sensors' data. Since people usually have different schedules between the weekdays and the weekend, we have scrutinised weekday household routines and weekend household routines separately.

The K-Means algorithm was applied to the motion sensor data and smart water meter data to perform cluster analysis. The objective was to group the data points in a meaningful way based on their similarities and extract valuable knowledge about the household routines and behaviours. We experimented with various values of K (number of clusters) ranging from 3 to 9 using the Elbow and Silhouette metrics and also our visual observations of the data towards acquiring the indispensable insights from the data. Based on this experimentation, we arrived to the conclusion that five clusters are appropriate for all the cases, since we observed that suchlike clustering can uncover useful and meaningful information regarding the routines of the household occupants, as for example the time they wake up in the morning, if they wake up during the night in order to use the toilet a lot of times indicating a possible health issue, if they wake up during the night and wander around the house possibly indicating sleep problems, stress, or mental health conditions, the times when they are away from the house (occupancy detection), etc. In this context, K-Means demonstrated its effectiveness in terms of both similarity within clusters and distance between points of different clusters. By minimising the within-cluster sum of squares, K-Means successfully grouped similar data points together, ensuring that the data points within each cluster share common characteristics or patterns. This supported the identification of meaningful clusters that captured specific behaviours, such as wake-up times, nighttime activities, wandering behaviours, and occupancy patterns. Furthermore, K-Means achieved separation between points of different clusters by considering the distance between data points. It assigned each data point to the cluster with the closest mean value based on the squared Euclidean distance. As a result, data points that were closer to each other in terms of their feature values were more likely to be assigned to the same cluster, while points that were farther apart were more likely to belong to different clusters. This ensured that points in different clusters were dissimilar to each other, enhancing the distinctiveness among the identified groups. Therefore, based on the experiments, it can be concluded that K-Means is effective in promoting similarity within clusters and achieving separation between points of different clusters, while it successfully picked up meaningful patterns within the data, allowing for valuable insights to be extracted about the household routines.

In the implementation of the DBSCAN models, the algorithm's parameters, epsilon and minimum samples, were tuned to find the optimal combination. However, due to the presence of variable density clusters in the data, the DBSCAN algorithm did not yield satisfactory results. Despite trying various combinations of epsilon and minimum samples values, it was challenging to find a configuration that could effectively capture all the clusters in a meaningful way. The DBSCAN algorithm relies on the density of clusters to identify and group data

Alexia Dini

points. However, in datasets with variable densities, such as the ones used in this study, the algorithm struggles to accurately detect all the clusters. Consequently, some clusters were missed or fragmented, making it difficult to obtain a comprehensive and meaningful clustering solution. The limitation of DBSCAN in handling datasets with variable densities is well-documented [79]. The algorithm's performance tends to degrade when faced with such data distributions. In this case, the dataset's density fluctuations posed a significant challenge in finding an appropriate set of epsilon and minimum samples values that could effectively capture all the clusters in a satisfactory manner. Therefore, despite the attempts to fine-tune the parameters and explore different combinations, the DBSCAN algorithm was not able to produce desirable results for the given datasets. Its reliance on density-based clustering hindered its ability to accurately identify and group the data points into meaningful clusters.

Finally, in the Agglomerative clustering algorithm, dendrograms were used to split clusters into numerous clusters of related data points. Different threshold values were used defining the minimum distance required to separate a cluster for splitting up the dendrograms. Nevertheless, the results produced by this model were not very effective in all the experiments in determining the correct number of clusters by the dendrogram and did not scale good enough according to the number of the data points. This can be attributed to the characteristics of the datasets used in the experiments. The smart water meter dataset might contain irregular patterns and fluctuations in water consumption, leading to diverse and non-uniform clusters. As Agglomerative clustering relies on proximity-based merging and splitting, it may struggle to identify the appropriate number of clusters based on the dendrogram. The irregularities and variations in the smart water meter dataset can hinder the algorithm's ability to accurately capture the underlying structure and determine the optimal number of clusters. Furthermore, the motion sensor dataset includes time-series data, collected at 10-second intervals. However, the Agglomerative clustering algorithm may have difficulties in handling time-series data and its scalability with increasing data points. As the number of data points in the motion sensor dataset grows, the computational complexity of the Agglomerative clustering increases, potentially leading to performance degradation. The algorithm's hierarchical merging and splitting process may become computationally expensive and less efficient for larger datasets. Consequently, this can impact its ability to determine the correct number of clusters and affect its overall scalability.

As an indication of the experimentation, the results of the three algorithms for the Smart Water Meter data using the NAZ approach can be seen in Figure 4.4, Figure 4.5, and Figure 4.6. Based on the comparison of the results from the three algorithms implemented,

Alexia Dini

we selected the K-Means algorithm as the best candidate for our study, as it produced good interpretable clusters that can enable the creation of rules for the data and the extraction of inferences. Moreover, K-Means is stable for time-series data and is suitable for huge datasets [174].

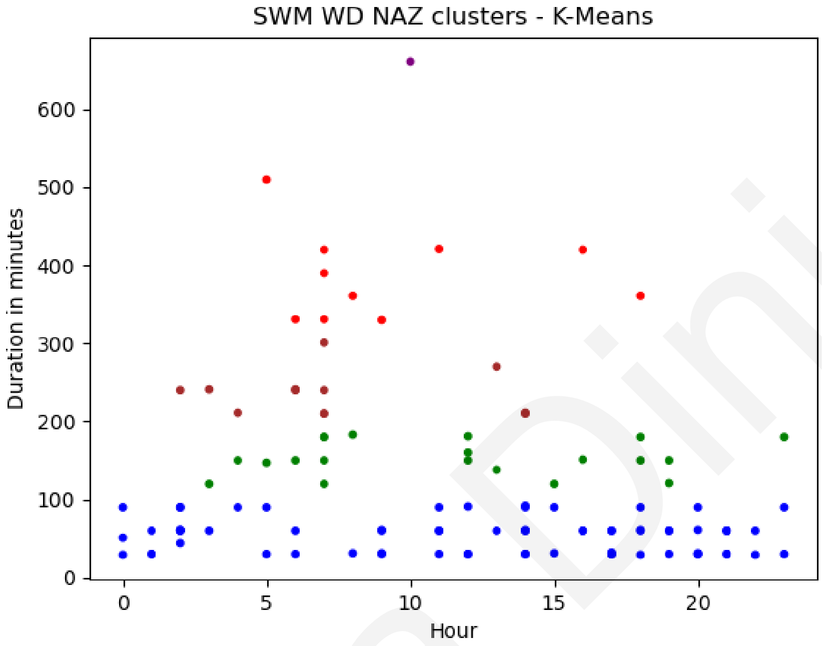


Figure 4.4: Results of the K-means Algorithm for the Smart Water Meter data using the NAZ approach

Alexia Dini

4.4.2 Rules definition and cluster interpretation

The next step was to train the K-Means model with the available dataset for each experiment. Then we created a Decision Tree model, which was trained using the original data features and the K-Means model clustering result as the label for each experiment in order to get an initial interpretation of what each cluster represents in the three experiments. Using two adjustable parameters, the min_samples_leaf, which is the minimum number of samples required to be at a leaf node, and the pruning_level, a technique that removes parts of the Decision Tree which prevent it from growing to its full depth, avoiding overfitting of the training data, we controlled the complexity of the Decision Tree. For the creation of more detailed rules we decreased one of those values accordingly. The decision tree leaves provided useful information that assisted in the specification of a number of conditions that the data tested satisfies. In each experiment, the relevant dataset was fitted to the decision tree and the results from the decision tree were compared with the data in the clusters in order to identify which conditions are satisfied by the data in each cluster. This served as a guidance

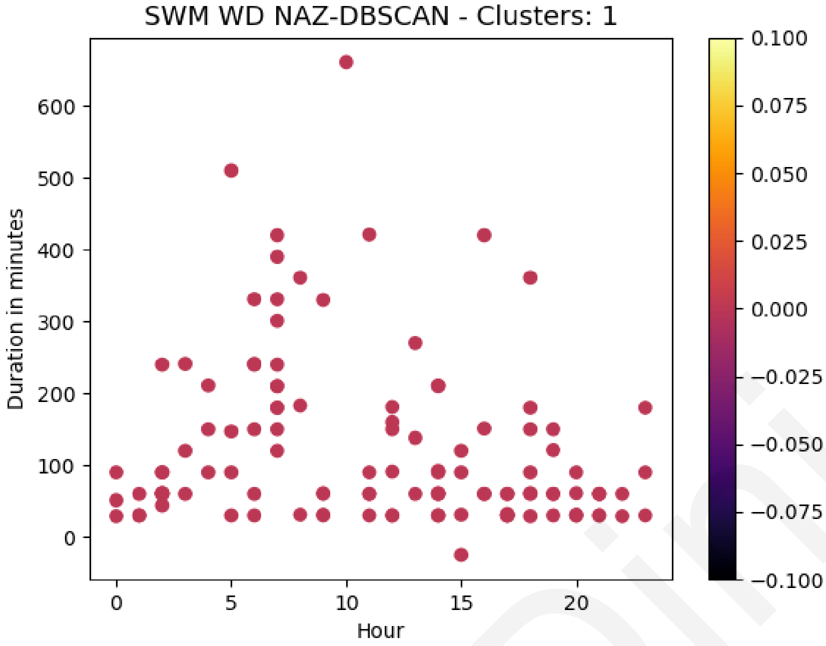


Figure 4.5: Results of the DBSCAN Algorithm for the Smart Water Meter data using the NAZ approach

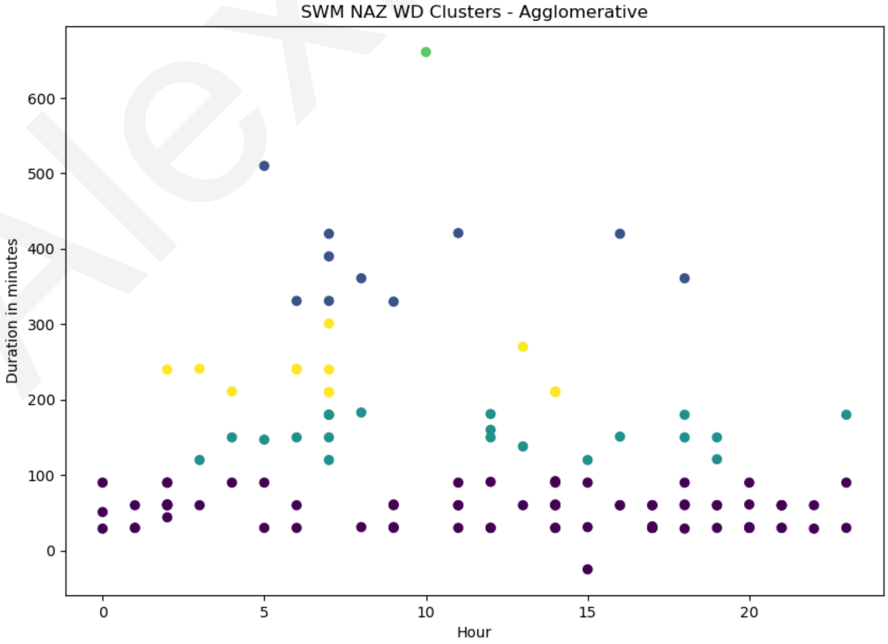
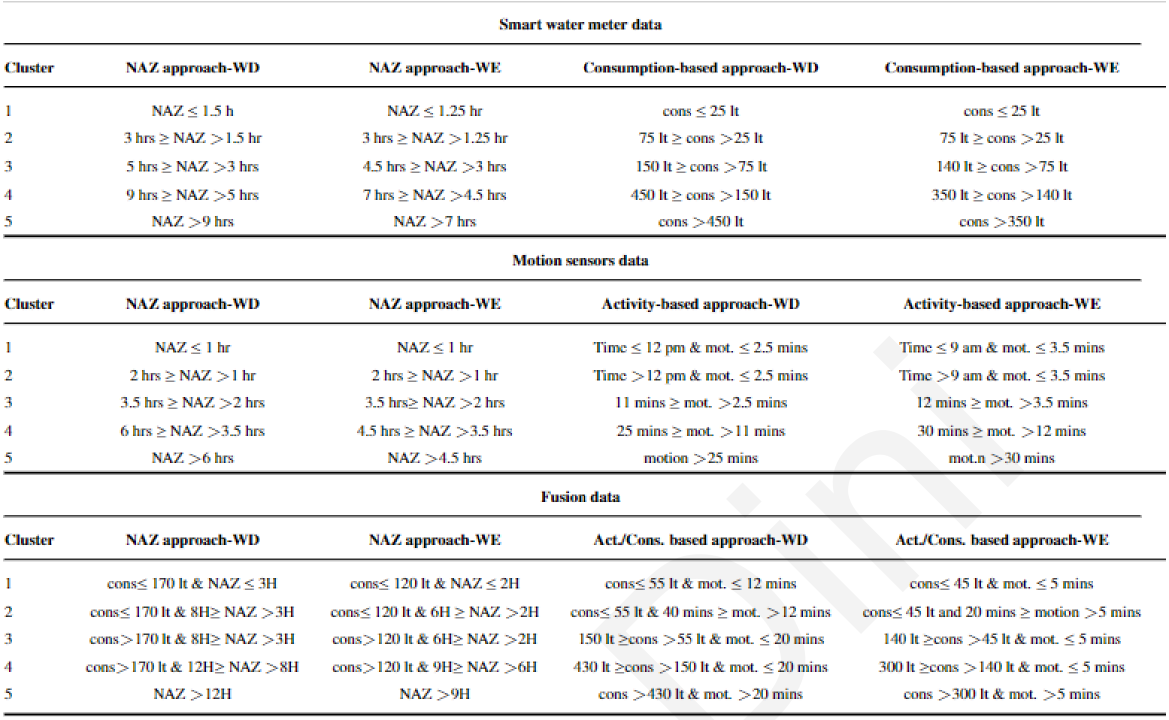


Figure 4.6: Results of the Agglomerative Hierarchical Clustering Algorithm for the Smart Water Meter data using the NAZ approach

Alexia Dini

for the generation of a list of rules for each possible approach/dataset combination. The list of rules derived in the smart home scenario experiments can be seen in Figure 4.7.



Smart water meter data

Cluster	NAZ approach-WD	NAZ approach-WE	Consumption-based approach-WD	Consumption-based approach-WE
1	NAZ ≤ 1.5 h	NAZ ≤ 1.25 hr	cons ≤ 25 lt	cons ≤ 25 lt
2	3 hrs ≥ NAZ >1.5 hr	3 hrs ≥ NAZ >1.25 hr	75 lt ≥ cons >25 lt	75 lt ≥ cons >25 lt
3	5 hrs ≥ NAZ >3 hrs	4.5 hrs ≥ NAZ >3 hrs	150 lt ≥ cons >75 lt	140 lt ≥ cons >75 lt
4	9 hrs ≥ NAZ >5 hrs	7 hrs ≥ NAZ >4.5 hrs	450 lt ≥ cons >150 lt	350 lt ≥ cons >140 lt
5	NAZ >9 hrs	NAZ >7 hrs	cons >450 lt	cons >350 lt

Motion sensors data

Cluster	NAZ approach-WD	NAZ approach-WE	Activity-based approach-WD	Activity-based approach-WE
1	NAZ ≤ 1 hr	NAZ ≤ 1 hr	Time ≤ 12 pm & mot. ≤ 2.5 mins	Time ≤ 9 am & mot. ≤ 3.5 mins
2	2 hrs ≥ NAZ >1 hr	2 hrs ≥ NAZ >1 hr	Time >12 pm & mot. ≤ 2.5 mins	Time >9 am & mot. ≤ 3.5 mins
3	3.5 hrs ≥ NAZ >2 hrs	3.5 hrs≥ NAZ >2 hrs	11 mins ≥ mot. >2.5 mins	12 mins ≥ mot. >3.5 mins
4	6 hrs ≥ NAZ >3.5 hrs	4.5 hrs ≥ NAZ >3.5 hrs	25 mins ≥ mot. >11 mins	30 mins ≥ mot. >12 mins
5	NAZ >6 hrs	NAZ >4.5 hrs	motion >25 mins	mot.n >30 mins

Fusion data

Cluster	NAZ approach-WD	NAZ approach-WE	Act./Cons. based approach-WD	Act./Cons. based approach-WE
1	cons≤ 170 lt & NAZ ≤ 3H	cons≤ 120 lt & NAZ ≤ 2H	cons≤ 55 lt & mot. ≤ 12 mins	cons≤ 45 lt & mot. ≤ 5 mins
2	cons≤ 170 lt & 8H≥ NAZ >3H	cons≤ 120 lt & 6H ≥ NAZ >2H	cons≤ 55 lt & 40 mins ≥ mot. >12 mins	cons≤ 45 lt and 20 mins ≥ motion >5 mins
3	cons>170 lt & 8H≥ NAZ >3H	cons>120 lt & 6H≥ NAZ >2H	150 lt ≥cons >55 lt & mot. ≤ 20 mins	140 lt ≥cons >45 lt & mot. ≤ 5 mins
4	cons>170 lt & 12H≥ NAZ >8H	cons>120 lt & 9H≥ NAZ >6H	430 lt ≥cons >150 lt & mot. ≤ 20 mins	300 lt ≥cons >140 lt & mot. ≤ 5 mins
5	NAZ >12H	NAZ >9H	cons >430 lt & mot. >20 mins	cons >300 lt & mot. >5 mins

Figure 4.7: List of rules derived in the smart home scenario experiments

Alexia Dini

4.4.3 Extracting Inferences

The next step in the experiments was the inference detection process. In each experiment, the clustering results of the two approaches were analyzed in order to determine what kind of information can be acquired that can provide answers to the questions set earlier. Consequently, the cluster that contained information relevant to what was required was located and in combination with the list of rules, a number of functions was applied that identified that information. The intention was to verify if and which of the identified inferences (Figure 4.2) can be determined through the sensors' data. Since people usually have different schedules between the weekdays and the weekend, weekday household routines and weekend household routines were scrutinised separately.

In the first experiment using the motion sensor dataset, the process disclosed information about the usual wake up time of the residents, the time of sleep, the time the residents leave for work or school in the morning, the time they return and how many times they get out of bed and wander in the house during the night. This was achieved by exploiting the recorded motion data during the day. Similar results were produced in the second experiment, where the smart water meter data was utilised for this purpose. In this case, sudden intensive water use in the morning indicated that the residents have woken up, scarce adequate water consumption during the night designated probable toilet use, while heavy water consumption during the day marked occupancy. For the identification of the water consumption levels for various household activities we used the information provided in [35]. An example of the results from smart water meter data experiment using the consumption-based approach can be seen in table 4.1.

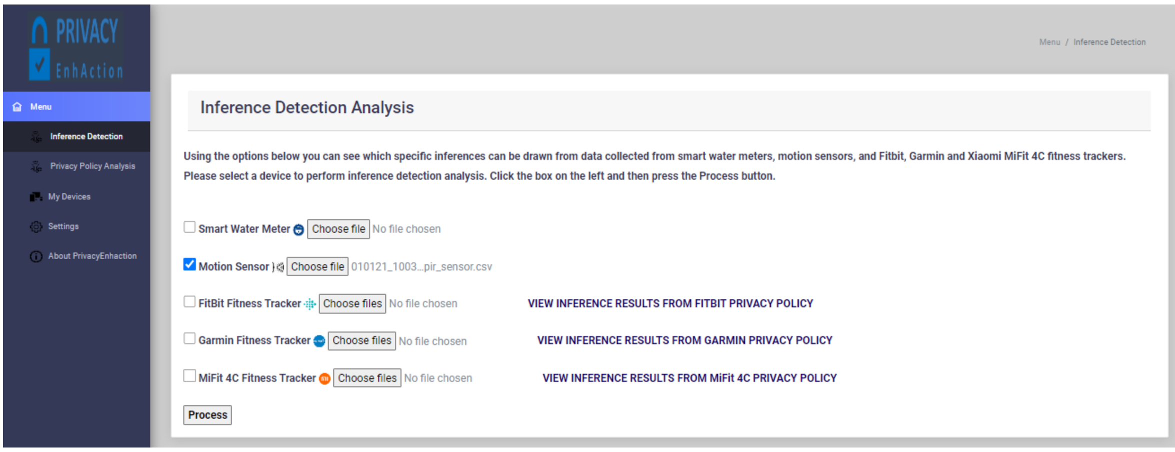
In the third experiment we combined the data from both the sensors. The additional information derived from the linkage of motion and water consumption data was the identification of the times when motion and water consumption was concurrent or non-existent, which can be used for further inferences. For example, we were able to identify if the user wakes up during the night to use the toilet. If this is something that happens a lot during the night, then this could probably imply health problems.

Table 4.1: Results for Smart Water Meter data using the consumption-based approach

Inference	**Weekday**	**Weekend**
Wake up time	06.30 a.m.	08:00 a.m.
Sleep time	22:30 p.m.	23:30 p.m.
No of times the users wake up at night	4 times	4 times
Departure time in the morning	08:00 a.m.	11:30 a.m.
Return time	14:00 p.m.	17:00 p.m.

Both approaches used in the realisation of the experiments produced approximately the same results. In the third experiment however, where we combine the data from the motion sensor and the smart water meter, the NAZ approach proved to be a better fit as we were able to deduct the exact start and end time of each NAZ, while when using the other approach we missed out on the actual times since hourly resampling was performed. Therefore, the use of the NAZ approach is recommended, as it can provide accurate start and end time of activities. In the effort to extract possible inferences from the combination of the available datasets, over and above the privacy inferences amassed, the experiments led to useful findings that can enhance the safety of the residences, such as the identification of water leakages during long-term absence from the house.

Alexia Dini


PRIVACY
EnhAction
Menu
Inference Detection
Privacy Policy Analysis
My Devices
Settings
About PrivacyEnhaction
Menu / Inference Detection
Choose file
Choose files
VIEW INFERENCE RESULTS FROM FITBIT PRIVACY POLICY
VIEW INFERENCE RESULTS FROM GARMIN PRIVACY POLICY
VIEW INFERENCE RESULTS FROM MiFit 4C PRIVACY POLICY
Process

Inference Detection Analysis

Using the options below you can see which specific inferences can be drawn from data collected from smart water meters, motion sensors, and Fitbit, Garmin and Xiaomi MiFit 4C fitness trackers. Please select a device to perform inference detection analysis. Click the box on the left and then press the Process button.

□ Smart Water Meter — No file chosen

☑ Motion Sensor — 010121_1003...pir_sensor.csv

□ FitBit Fitness Tracker — No file chosen

□ Garmin Fitness Tracker — No file chosen

□ MiFit 4C Fitness Tracker — No file chosen

Figure 4.8: Initial PrivacyEnhAction Inference Detection page

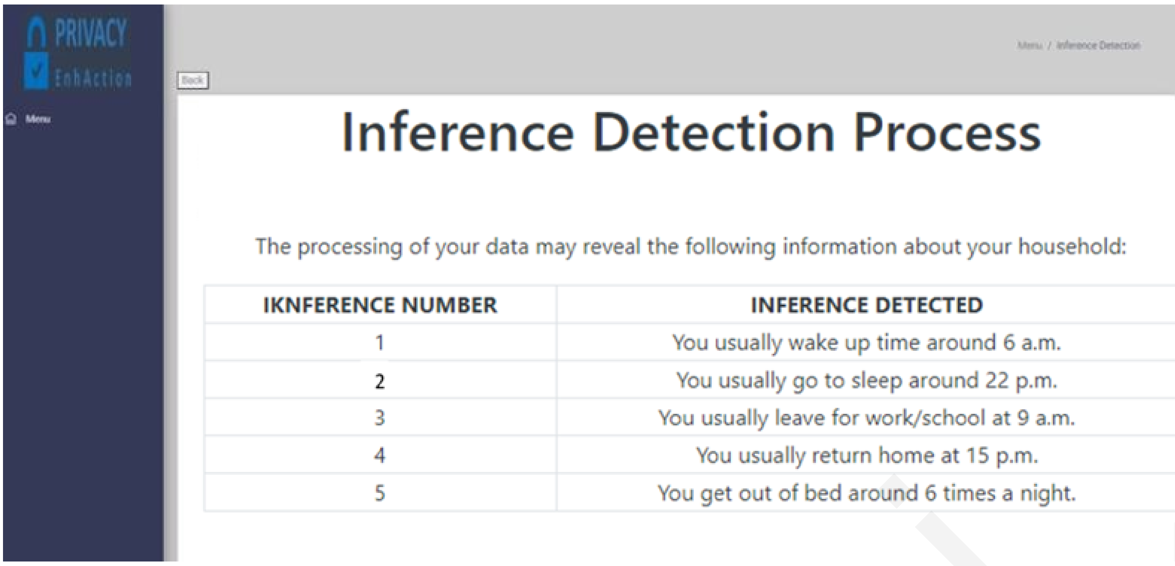
4.5 Proof of concept application

In order to evaluate our models, we used the web application that we have created for this purpose, PrivacyEnhAction. This application is a privacy tool that can be used by the owners of smart home devices and fitness trackers in order to analyse the data generated by their devices and be informed about what possible inferences can be drawn from these data that may violate their privacy. The tool was created using the Flask micro-framework and all the code was written in Python.

Since the experiments were performed using the specific smart home scenario, the models were trained on the datasets generated by the specific household smart devices. In order to get feedback from users, we performed a small scale user evaluation of the tool with 5 users. We provided a number of datasets originating from the smart home setting to a small number of people that were willing to use and evaluate the platform in terms of usability and feasibility. In Figure 4.8 the Inference Detection page of the tool is shown where the user has chosen to check the Motion Sensor data for inferences, while Figure 4.9 shows the detected inferences for the supplied data.

Through our interaction with the users, we collected feedback about user experience and user satisfaction. Overall, the users thought that the platform is easy to use. They were satisfied with the results and found the information received very explanatory. What one of the users commented as an additional useful feature is the possibility to see if data collection can be modified, which is something we envision as part of future work.

Alexia Dini



Inference Detection Process

The processing of your data may reveal the following information about your household:

IKNFERENCE NUMBER	INFERENCE DETECTED
1	You usually wake up time around 6 a.m.
2	You usually go to sleep around 22 p.m.
3	You usually leave for work/school at 9 a.m.
4	You usually return home at 15 p.m.
5	You get out of bed around 6 times a night.

Figure 4.9: PrivacyEnhAction results for Smart Motion sensor data

4.6 Discussion

In this Chapter we investigated the possibility of extracting inferences about user routines from smart home data, providing an answer to Research Question **RQ2**: ***"What inferences can be made from data collected from smart home devices and fitness trackers?"***. We analysed smart water meter and motion sensors data in order to identify what these inferences may be and notify the users about the privacy risks they face when using such devices. For this purpose, we implemented a privacy tool, PrivacyEnhAction, a web application that can be used by smart home devices and fitness trackers users in order to perform an analysis on the data created by their devices and be informed about the kind of information that can be obtained about them. This knowledge will empower the users to choose whether they wish to continue using a particular device, or to make amendments in the privacy settings of the device, if available, altering for example the frequency of data collection.

The models developed for the PrivacyEnhAction tool and presented in this Chapter have been trained according to the specific household's routines. Still, they can easily be adapted for other users as well by getting their feedback regarding their own daily activities, and retrain the models. The methodology used can be adapted in other scenarios as well as it is not bound to smart home scenarios. Through our experiments we have showed that the models are viable and can be used as a tool for enhancing the protection of the users' privacy in a smart home setting. Moreover, we trust that our approach would be beneficial to IoT service providers or developers of smart devices serving as a guideline for the protection of privacy and the provision of better services to the users.

Alexia Dini

Chapter 5

Inference detection in the fitness trackers domain

Alexia Dini

In this Chapter, we concentrate on fitness trackers from three brands, namely Fitbit, Garmin and Xiaomi, in order to investigate if the analysis and exploitation of the data collected by those trackers can lead to the extraction of inferences about the owners routines, health status or other sensitive information, providing an answer to Research Question **RQ2**: ***“What inferences can be made from data collected from smart home devices and fitness trackers?"***, in order to inform the users about the findings concerning data inferences through the PrivacyEnhAction web application. We utilise the data inference framework introduced in Chapter 4, where by using a number of statistical analysis and modelling techniques we aim to verify that such inferences are possible in order to raise user awareness about them. These techniques are implemented in the PrivacyEnhAction privacy tool, through which the users can analyse data collected from their smart home devices or fitness trackers with the objective to be informed about potential privacy vulnerabilities and possible inferences that emerge from the use of these devices, and thereupon to be able to change and set their user privacy preferences on their devices appropriately, contributing in this way to the personalisation of the provided services, in connection with their personal data. To that end, our work is user-oriented aiming to raise user awareness regarding privacy in the area of IoT. The work presented in this chapter is based on our research published in the Springer *User Modeling and User-Adapted Interaction* journal [83].

5.1 Possible inferences from fitness trackers data

What we aim to address in this chapter is to provide awareness to the users about the possible privacy risks and the inferences that can be extracted about them from their fitness trackers data, so that they can set their user privacy preferences in such a way that their personal privacy can be protected. By accomplishing this task, we also aim to provide an answer to Research Question **RQ2**: ***"What inferences can be made from data collected from smart home devices and fitness trackers?"***. In order to answer this question we use the results from the literature review we performed for this work in combination with our previous research in the thesis, and we produce a list of possible inferences that pose a threat to user privacy when using fitness trackers. We also aim to find which inferences can be drawn from the data collected from the specific fitness trackers in this study. The derived list of possible inferences that form a threat to user privacy when using fitness trackers can be seen in Table 5.1.

5.2 Fitness trackers scenarios under study

A big number of commercial fitness tracker devices are available on the market from different manufacturers, an indicative list of which can be seen in Table 5.2. For the purposes of this work, we have chosen to employ Fitbit and Garmin fitness trackers after reviewing the available literature, where Fitbit and Garmin devices were identified as the most popular devices [279]. Moreover, Fitbit Surge and Garmin Forerunner appear to have embedded the biggest number of sensors, i.e., PPG, GPS, gyroscope, magnetometer, and barometer or altimeter [128], which means that these devices collect more user data. We have also chosen to include Xiaomi fitness trackers in our study, as Xiaomi appeared in the top five vendors in sales for two consecutive years (2015 and 2016) [128] and also due to their low cost as our budget was limited.

After the possible inferences that can be extracted from fitness trackers data have been identified, the next step is to find which inferences can be drawn from the data collected from the specific fitness trackers in this study. We also describe the methodology we used in this study in order to collect, examine and analyse the data in the fitness trackers scenarios, following the methodology we proposed in Chapter 4, adjusted to suit the current study's needs, which can be applied in other IoT scenarios with minor modifications.

5.2.1 Data collection process

In this section, we provide details about the data collection process, in relevance to how we gathered our participants and what mechanisms we used for the data collection.

Alexia Dini

Table 5.1: Possible fitness trackers inferences

Data Type	Inference	Interested 3rd Party	Potential Use	An/sis	Sam.Size	Study
Activity data						
Phys. activity	Health problems	Ins. companies	Increased rates	T	n/a	[170]
Phys. activity	Chronic diseases	Ins. companies	Increased rates	T	n/a	[38]
Phys. activity	Mortality risk	Ins. companies	Increased rates	T	n/a	[38]
Phys. activity	Human emotions	Employer	Discrimination	T	n/a	[164]
Activity, location	Religion	Employer	Discrimination	E	970	[143]
Activity, location	Religion	Employer	Discrimination	E	227	[302]
VO2max	Fitness level	Ins. companies	Increased rates	E	10	[311]
Heart rate data						
Resting heart rate	Pregnancy likelihood	Employer	Discrimination	E	8	[5]
Heart rate, respiration	Substance abuse	Ins. companies, employer	Discr., incr. rates	T	n/a	[224]
Heart rate, accelerometer data	Sexual activity	Marketing companies	Targeted advertising	E	227	[302]
High resting heart rate	Health problems, alcohol abuse	Ins. companies, employer	Discr., incr. rates	E	21853	[68]
High resting heart rate	Health problems, alcohol abuse	Ins. companies, employer	Discr., incr. rates	E	6743	[10]
Low resting heart rate	Bradycardia, medication	Ins. companies, employer	Discr., incr. rates	T	n/a	[188]
GPS data						
GPS data	Location tracking	Attackers	Targeted home or personal attacks	T	n/a	[18]
GPS data	Frequently visited places	Attackers	Personal attacks	E	n/a	[196]
GPS data	Location	Marketing companies	Targeted advertising	T	n/a	[105]
GPS data	Health, habits, prof. duties	Attackers	User profiling	T	n/a	[71]
GPS data	Location	Many	Political, religious, sexual discrimination	T	n/a	[71]
Sleep data						
Sleep data	Sleep deprecation	Ins. companies, employer	Discr., incr. rates	T	n/a	[131]
Sleep data	Sleep patterns	Marketing companies	Targeted advertising	T	n/a	[299]
User weight,height	Obesity	Insurance companies	Increased rates	E	n/a	[330]

Abbr. for Analysis: T=Theoretical; E=Experimental **Abbr. for Sample Study:** n/a = Not Applicable

Alexia Dini

Table 5.2: Indicative list of available commercial fitness trackers

Manufacturer	**FT Models**
Fitbit	Surge, Charge, Ace, Inspire, Luxe
Xiaomi	Mi Smart Band 4C, 5, 6, 7,Redmi Watch 2, Redmi Smart Band Pro
Garmin	Forerunner, Captain, Fenix, Epix, Venu, Vivosmart, Vivofit, Instict, Quatix
Apple	Apple Watch
Huawei	Band 6, Band 4 Pro, Band 4, Band 4e
Amazifit	Band 7, Band 5, Verge, Nexo, X
Samsung	Galaxy Watch 4
Withings	Scan Watch, Steel HR
Polar	Grit X Pro
Suunto	Peak, Baro

Alexia Dini

Participant recruitment

We recruited participants by sending email invitations to members of the SEIT Lab[1] of the University of Cyprus. In total, 5 people responded who were fit to participate in the study, meaning that they were over 18 years old and were not diagnosed with any chronic disease. As more participants were required, family and friends of the authors were recruited that fit the criteria. All participants provided their informed consent for submitting their personal data. The details of the participants can be found in Table 5.3. Before the data collection period started, a meeting was held with the participants in order to inform them about what was required from them, to assist them with setting up the necessary environment by installing the required apps on their mobile phones and to create personal accounts for the devices.

Data collection mechanisms

For the collection of data, we have acquired one FitBit Surge fitness tracker, five Xiaomi Mi Smart Band 4C devices and two Garmin smart watches, that were assigned to eight participants respectively, who were asked to wear them for 24 hours a day for a period of two months. As more data were necessary for our experiments, we explored various online repositories, such as Zenodo and Kaggle, in order to find additional fitness tracker datasets.

[1] https://www.cs.ucy.ac.cy/seit/

Table 5.3: Participants demographics

User	Gender	Age	FT Model and Brand
1	Female	45	Fitbit Surge
2	Male	26	Mi Smart Band 4C
3	Female	28	Mi Smart Band 4C
4	Female	45	Mi Smart Band 4C
5	Female	38	Mi Smart Band 4C
6	Female	20	Mi Smart Band 4C
7	Female	38	Garmin Forerunner 630
8	Male	48	Garmin Captain

Table 5.4: Fitness tracker datasets downloaded from repositories

Brand	Dataset	Repository
Fitbit	Crowd-sourced Fitbit datasets[2]	Zenodo
Garmin	Run Activities[3]	Kaggle
Mi Band	5 years of continuous steps and sleep data[4]	Kaggle
Mi Band	Exported data from Mi Band fitness tracker[5]	Kaggle

Alexia Dini

Due to the sensitive nature of the data involved, finding suitable public datasets was not an easy task. Still, we located a small number of fitness tracker datasets suitable for our experiments, more details of which can be seen in Table 5.4.

5.2.2 Data processing and cleaning

In this section, we provide information about how the available datasets were processed and cleaned in order to be ready for the next step of data analysis.

Fitbit datasets

For the first experiment we employed a Fitbit Surge device owned by one of the participants and we also used the public dataset “Crowd-sourced Fitbit datasets" available at the Zenodo

[2] https://doi.org/10.5281/zenodo.53894

[3] https://www.kaggle.com/mmaelicke/run-activites

[4] https://www.kaggle.com/damirgadylyaev/more-than-4-years-of-steps-and-sleep-data-mi-band

[5] https://www.kaggle.com/bekbolsky/exported-data-from-xiaomi-mi-band-fitness-tracker

repository [107]. This dataset was collected by thirty eligible Fitbit users that participated in an Amazon Mechanical Turk survey, submitting physical activity, heart rate, and sleep monitoring data at minute level. In this dataset, different types of data are stored in 18 files in total, where each file contains merged data from the different users. In order to derive suitable data for our experiment in separate sets for each user, we manually processed the dataset by parsing each file by export session ID that corresponds to a unique user. Following this procedure we acquired a number of user datasets, containing daily physical activity data, heart rate and sleep monitoring data. Each dataset represents a unique user and consists of three files in .csv format. Data processing also required deleting any records containing missing or null values and removing any outliers identified.

Alexia Dini

Garmin datasets

In this experiment, two volunteers were assigned to wear a Garmin smart watch for two months. Then each volunteer's data was exported through Garmin Connect using the Request Data Export option. The exported datasets consisted of a number of files in JavaScript Object Notation format (JSON), which were then converted to a CSV format using a JSON to CSV converter tool. Manual examination of the files' content assisted in determining which specific data would be useful for data analysis. This process resulted in the acquisition of two files in each dataset at this stage, the first containing general activity data like activity name, activity type, timestamp, duration, distance, calories, startLongitude, startLatitude, avgHr, maxHr, vO2MaxValue, etc. and the second containing sleep data. Again, data processing required deleting any records containing missing or null values and removing any outlier values identified.

Xiaomi datasets

For this experiment we acquired five Mi Smart Band 4C devices, that were allocated to five participants who wore them for 24 hours for a period of two months. When the data collection cycle ended, each participant's data was exported using the Mi Fit account 'Export Data' option. The datasets received consisted of a number of folders with data in CSV format, whose content was manually examined in order to evaluate which data would be suitable for analysis. This method led to the inclusion of four files in each dataset, containing activity data, heart rate data, sleep data and user information. Any records with null values or missing data were removed from the files.

5.2.3 Data analysis techniques

In order to analyse the data, we used statistical analysis and descriptive analytics techniques in our effort to assess and understand the available data. Using the fitness trackers datasets we had at our disposal, we performed Exploratory Data Analysis (EDA), aiming to identify patterns or anomalies on the data using summary statistics and graphical representations, with the intention to identify if any particular data points or the combination of them will facilitate the elicitation of one or more of the designated inferences. EDA is a method that uses data visualisation on datasets in order to determine the relationships of data aiming to find patterns that can reveal hidden information in the data [235]. Correlation analysis, an EDA technique used to measure the strength of the linear relationship between two variables [247], was applied in order to evaluate the relationships between variables, as any potential connection between variables can enable the extraction of useful information from the data.

Alexia Dini

5.3 Inference identification in fitness trackers under study

Based on the available data and in line with the analysis performed in the previous section, we undertook the task to identify which inferences can be extracted in accordance to the inferences list defined in Table 5.1. It must be noted that the inferences identified in this chapter are only indications and cannot be used as a verification or evidence. For example, if the available user resting heart rate data can lead to the conclusion that the female user may be pregnant, this inference is not a proof that the particular user is indeed pregnant, but it is only an indication that the user *may be* pregnant.

5.3.1 Fitbit inference detection analysis

Inferences from Fitbit heart rate data: Fitbit heart rate data contain heart rate measurements at 5-seconds intervals. According to Table 5.1, using the heart rate measurements we can try to infer: (a) pregnancy possibility, (b) whether the user suffers from health problems in general, (c) alcohol abuse, or (d) whether the user is under medication. The procedure described next was adopted for this purpose.

In order to infer pregnancy possibility, information about the user gender is necessary. As this piece of information was not included in the available Fitbit datasets, we did not attempt to extract this insight from the rest of the data, e.g. the resting heart rate.

An elevated or low resting heart rate can assist in extracting inferences (b), (c) and (d). From analysing the available datasets, no information about specific activity and activity times was given, that could be excluded from further analysis. It was then decided to utilise the available sleep data instead. To this extent, heart rate data was combined with sleep data to match sleeping times with corresponding heart rate values and thus extract the resting periods of the user. Using the new combined data, groups of heart rate measurements were created in the cases when there were successive values of above 100 beats per minute and a method was applied to the data to sum up the time between the minimum and the maximum timestamp of each group in order to find the length of time that the elevated heart rate lasted for. From this data it can be observed that when there are many long periods of time with elevated heart rate, then the inference that can be made is that the user may be suffering from health problems, since the heart rate is elevated during rest time (specifically sleep time). The same procedure was employed for finding the periods of time that the user had a low heart rate (below 60 beats per minute), and if there are many such periods, then it can be inferred that the user may be suffering from bradycardia or may be under medication.

The likelihood of user alcohol abuse can be inferred by using a combination of the available heart rate data and the sleep data, excluding heart rate measurements that fall within the sleeping range. The remaining heart rate data were utilised, creating groups of heart rate measurements when there were successive values above 100 beats per minute and applying a similar method as before for summing up the time between the minimum and the maximum timestamp of each group to find the length of time that the elevated heart rate lasted for. In particular, if the start and end times of these periods follow the same trend, for example at midnight near the time when the bars close, this could be an indication that the user could be an alcoholic.

Inferences from Fitbit activity data: From the Fitbit daily activity data, we can estimate the activity level of the user. In order to match the activity level of the user to the indices in Table 2.3, we proceeded by finding the value at which the variable for the Total Steps tended to cluster. Based on this value, we could infer the activity level of the user, and as a result whether the user leads a healthy lifestyle or not. Another inference we worked on using the available total steps data was the religion. Based on this, we calculated the average daily number of steps and we compared them against the average Saturday steps. If the difference between the two values implies that the Saturday activity is unusually low, then we have an indication (not a proof) that this person could be an observant Jew.

Inferences from Fitbit sleep data: Through an accelerometer and the LED located on the

Alexia Dini

back of the watch or fitness device, the Fitbit fitness tracker can detect when a user is sleeping and what stage of sleep he or she is in. In order to get insights from the available Fitbit sleep data, we calculated the start and end time of sleep for each calendar day in the sleep dataset. We also aggregated the total sleep time, as well as the total minutes in Light sleep, Deep sleep and REM sleep stages for each day, followed by the estimation of the values at which all these variables tend to cluster. We separated our calculations for weekday and weekend observations, as typically users are likely to have different habits between them. Following this process we could calculate approximately how many hours of sleep the user gets during the week and the weekend, the time that the user wakes up and goes to sleep, and the percentage of his sleep in light, deep and REM stages. Using this information, we can get an insight on whether the user gets enough sleep and her sleep patterns.

Alexia Dini

5.3.2 Garmin inference detection analysis

Inferences from Garmin activity data: Garmin activity data contain detailed information about user activities, such as running, cycling, etc. Using these data we were able to extract insights regarding the user's most frequent activities, and then exploiting the available information about the geographic coordinates (latitude and longitude) of the activity, we applied a reverse geo-coding process in order to find the places that the user's most usual activities take place. Garmin activity data also contain VO2max measurements, which we exploited over time in order to determine if the specific user has increased or decreased her fitness level. Based on these findings it can then be inferred whether the user is an athlete, and also her overall health status, as the variations of the VO2max values are widely used as an indicator of health.

Inferences from Garmin sleep data: Many Garmin devices have an optical heart rate sensor that utilises an Advanced Sleep Monitoring (ASM) feature, with which users have the ability to track their sleep statistics when wearing the watch while sleeping. Advanced sleep tracking is cut out for recognizing when the user falls asleep, wakes up, as well as acknowledging the sleep stages of the user throughout the night. Sleep stages include light, deep and REM sleep, which are determined by merging the heart rate, heart rate variability, respiration rate, body movement and other measurements. In our analysis of the available Garmin sleep data, we proceeded by calculating first the total sleep time for each night in the dataset and we determined the regularity of the weekly and weekend sleeping habits of the user. We also aggregated the total minutes in Light sleep, Deep sleep, and REM sleep stages for each night, the total awake minutes of each night, followed by the estimation of the values at which all these variables tend to cluster. We separated our calculations for weekday and weekend observations, as typically users are likely to have different habits during the week and the weekend. Following this process we could infer approximately how many hours of sleep the user gets during the week and the weekend, together with the time the user goes to sleep and the time she wakes up. Similar as before, this information can reveal if the user experiences sleep issues like lack of sleep, and if such information is shared with third parties, such as a current or potential employer, then the user may face unfair dismissal or employment discrimination. Using the inferred data about the average percentage of light sleep, deep sleep and REM sleep stages, one can draw conclusions regarding the user focus ability, her mood or memory, that the user is possibly under medications like antidepressants, that she may be suffering from anxiety or depression, among others.

5.3.3 Mi Fit inference detection analysis

Inferences from Mi Fit activity data: Mi Fit fitness trackers track activities like walking or running, number of steps taken, etc. Using the available Mi Fit activity data, the daily number of steps was exploited in order to estimate the activity level of the user. An analysis on the data was performed and then the value at which the steps variable tends to cluster was determined. Based on this value and the activity indices in Table 2.3, the activity level of the user could be determined and therefore whether the user leads a healthy lifestyle or not. The number of daily total steps was exploited in this scenario for the religion inference discussed in Section 5.3.1, where we followed the same approach in order to calculate the average daily number of steps and then compared this value against the average number of steps taken on Saturdays. If the difference between the two values implies that the Saturday activity is unusually low, then there is a likelihood that this person could be an observant Jew.

Inferences from Mi Fit heart rate data: The Xiaomi Mi Band collects heart rate measurements at regular intervals set by the user. We followed the same procedure as in the Fitbit heart rate data analysis in Section 5.3.1, and we managed to infer whether the user suffers from health problems in general, alcohol abuse, and whether the user is under medication. More user information was available in the Xiaomi datasets, including gender details, therefore we attempted to use these data to infer pregnancy likelihood. Resting heart rate measurements can be used in combination with the gender to infer pregnancy possibility. Considering

Alexia Dini

that the resting heart rate increases by 30-50% during pregnancy to match the needs of the growing baby [186], [137], we exploited the available personal user information in the Mi Band data that includes the user gender and date of birth, in order to infer the likelihood of pregnancy. We proceeded by combining the available sleep, user and heart rate data, with a view to isolate the data enclosing resting heart rate measurements and upon that we applied a test in order to check if these values fall in the increased by 30-50% range suggesting a possible pregnancy. Information about a person such as pregnancy could reveal information about that persons health and is classified as special category data in GDPR[6]. To that end, if this type of information is obtained by a third party, it can be used in a discriminatory way against that person.

Inferences from Mi Fit sleep data: The Mi Fit Band uses embedded sensors like accelerometer, gyroscope and PPG (heart rate monitor) to monitor user sleep by tracking body movements and heart rate. The band can also determine whether the user is in light sleep stage, deep sleep stage or REM sleep stage. We followed the same process we applied for analysing the Garmin sleep data, and we manage to infer the hours of sleep the user gets during the week and during the weekend, along with the time the user usually goes to sleep and wake up during the week and the weekend. Likewise the Fitbit and the Garmin scenarios, the information we extracted can disclose whether the user encounters any sleep problems like lack of sleep or insomnia. If such knowledge is shared with third parties, such as the user's current employer or a potential employer, then the user may face unfair dismissal or employment discrimination. In our analysis we also calculated the percentage of the user's sleep in light sleep stage, deep sleep stage and REM sleep stage, information that can be used to draw conclusions about the user's ability to focus throughout the day, her memory or mood. This information can also indicate that the user may take medications like antidepressants and that she may be suffering from anxiety or depression.

5.4 Implementation

The need for tools that will make the users aware of the privacy risks and the possible inferences that can be made about them from their fitness trackers data is now more important than ever, especially under the GDPR requirements. In Chapter 4 we introduced PrivacyEnhAction, the web application that aims to inform the users about potential privacy vulnerabilities

[6] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data

Alexia Dini

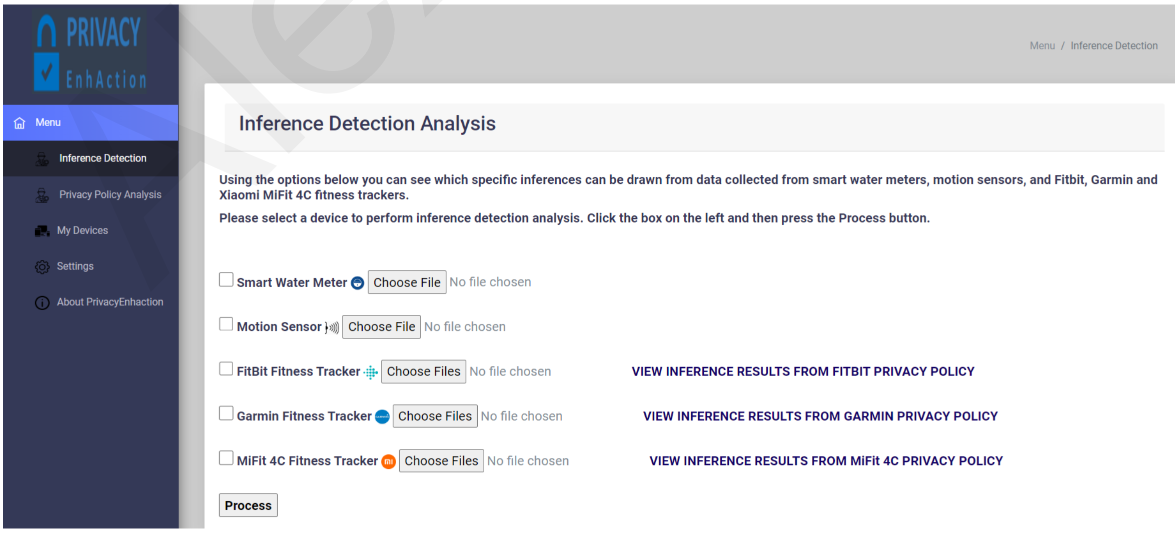
that emerge from the use of smart home devices and fitness trackers. In this section, we describe how we extended this tool by adding the three fitness trackers to the list of smart devices whose data can be analysed.

5.4.1 Extensions to the PrivacyEnhAction tool

PrivacyEnhAction has now been extended to include Fitbit, Garmin and MiFit 4C fitness trackers in the list of the available devices. The additional implemented functionalities consist of the following: (1) Inference detection analysis for Fitbit Surge; (2) Inference detection analysis for Garmin Captain and Garmin Forerunner 630 models; and (3) Inference detection analysis for Xiaomi Mi SmartBand 4C.

The code written in Python has been changed to include the new changes, where, depending on the selected device, the corresponding modules are called to process the files that are uploaded by the user. Using a number of statistical analysis methods, the application displays to the user the data-driven conclusions and possible inferences that can be drawn from her data. Then the user can select to view more information about each inference type, along with the possible risks that exist in relation to their privacy. Dedicated templates have been developed for each option that are rendered accordingly. The user interface has been adapted to reflect the new additions to the system following Nielsen and Molich's 10 user interface design guidelines [200] retaining all graphic representations and text across every system template.

Alexia Dini



Inference Detection Analysis

Using the options below you can see which specific inferences can be drawn from data collected from smart water meters, motion sensors, and Fitbit, Garmin and Xiaomi MiFit 4C fitness trackers.

Please select a device to perform inference detection analysis. Click the box on the left and then press the Process button.

☐ Smart Water Meter [Choose File] No file chosen

☐ Motion Sensor [Choose File] No file chosen

☐ FitBit Fitness Tracker [Choose Files] No file chosen — VIEW INFERENCE RESULTS FROM FITBIT PRIVACY POLICY

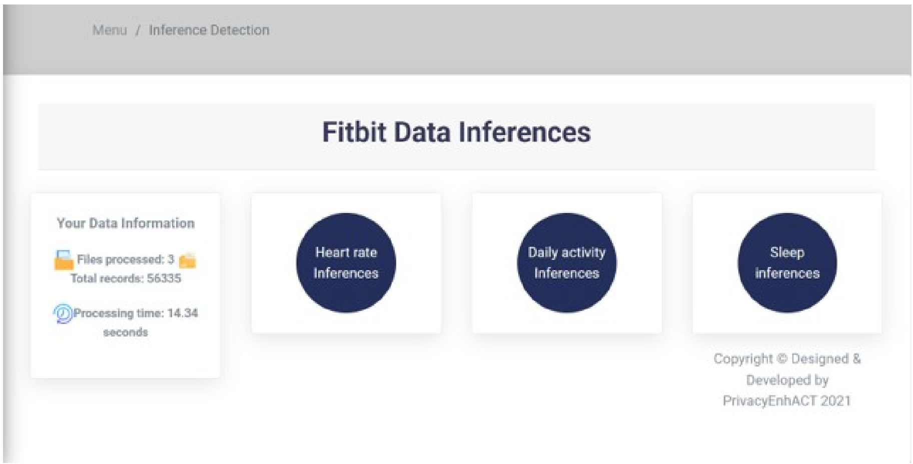
☐ Garmin Fitness Tracker [Choose Files] No file chosen — VIEW INFERENCE RESULTS FROM GARMIN PRIVACY POLICY

☐ MiFit 4C Fitness Tracker [Choose Files] No file chosen — VIEW INFERENCE RESULTS FROM MiFit 4C PRIVACY POLICY

[Process]

Figure 5.1: PrivacyEnhAction Inference Detection Analysis page

Users of these fitness trackers models can upload their data to the application through the interface, after they have exported them from their corresponding account dashboard, in



Fitbit Data Inferences

Your Data Information

Files processed: 3
Total records: 56335

Processing time: 14.34 seconds

Copyright © Designed & Developed by PrivacyEnhACT 2021

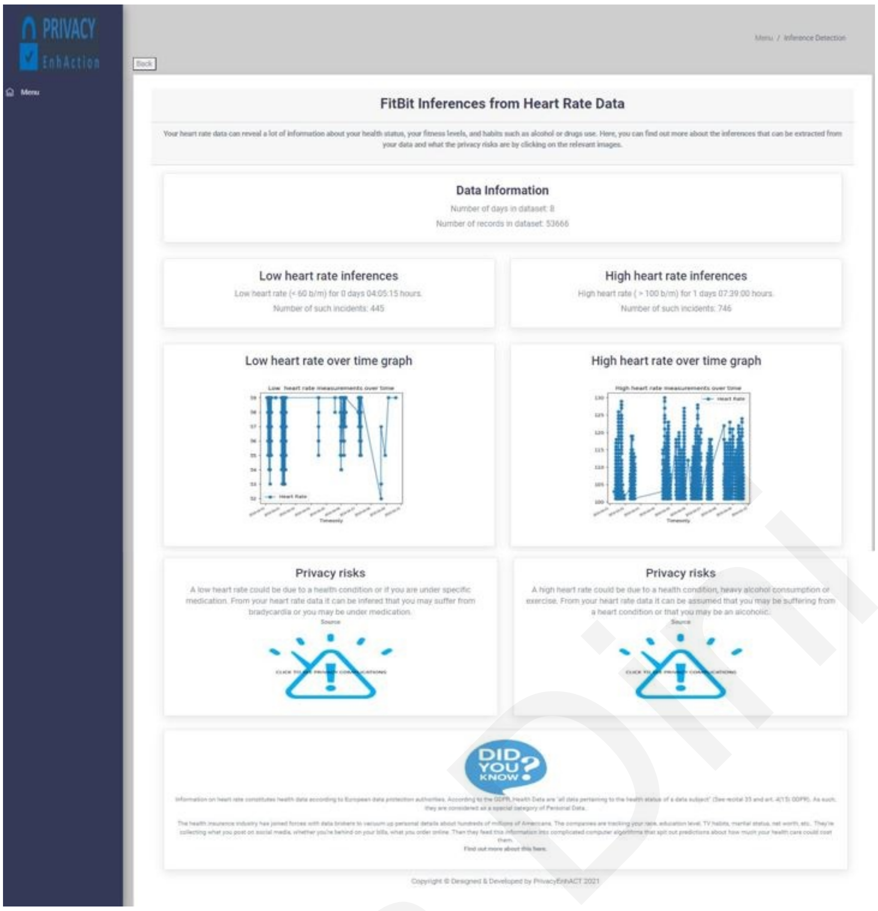
Figure 5.2: Screenshot from PrivacyEnhAction Fitbit data inferences results page

order to analyse the data and view the possible inferences that could be extracted about them and be informed about the potential privacy risks that these inferences entail. Figure 5.1 illustrates in a screenshot the Inference Detection Analysis page of the application, where the users can select the device they want to test for inferences.

When the data is processed, the application presents to the users the different types of inferences that could be drawn from their data, as illustrated in Figure 5.2, in which case the user has analysed Fitbit data. The privacy risks for each inference type are demonstrated by clicking on the corresponding button through the use of textual information and graphs related to the user's data, as well as further educational information, messages and links, as portrayed in Figure 5.3.

In the illustrated example regarding the inferences that can be extracted from the user's Fitbit heart rate data, the user is informed about what information can be revealed from the heart rate in the first part of the interface. Further down, the number of days and records in the processed dataset are displayed. In the next block, the user can view the number of incidents where her heart rate was below 60 bpm or over 100 bpm (low and high heart rate inferences respectively) during the days processed, as well as the total time that these events lasted for. The graphs of low heart rate and high heart rate over time for this time period are then presented. The user is then informed about the privacy risks and the insights that could be drawn about them from their heart rate data and by clicking on the blue sign we aim to increase the user's awareness by letting the user know how this information could be used by interested third parties. Inferences that could be obtained from Garmin and Mi Smart Band 4C fitness trackers are presented to the users in a similar manner.

Alexia Dini



FitBit Inferences from Heart Rate Data

Your heart rate data can reveal a lot of information about your health status, your fitness levels, and habits such as alcohol or drugs use. Here, you can find out more about the inferences that can be extracted from your data and what the privacy risks are by clicking on the relevant images.

Data Information

Number of days in dataset: 8

Number of records in dataset: 53666

Low heart rate inferences

Low heart rate (< 60 b/m) for 0 days 04:05:15 hours.

Number of such incidents: 445

High heart rate inferences

High heart rate (> 100 b/m) for 1 days 07:39:00 hours.

Number of such incidents: 746

Low heart rate over time graph

High heart rate over time graph

Privacy risks

A low heart rate could be due to a health condition or if you are under specific medication. From your heart rate data it can be infered that you may suffer from bradycardia or you may be under medication.

Source

Privacy risks

A high heart rate could be due to a health condition, heavy alcohol consumption or exercise. From your heart rate data it can be assumed that you may be suffering from a heart condition or that you may be an alcoholic.

Source

Information on heart rate constitutes health data according to European data protection authorities. According to the GDPR, Health Data are "all data pertaining to the health status of a data subject" (See recital 35 and art. 4(15) GDPR). As such, they are considered as a special category of Personal Data.

The health insurance industry has joined forces with data brokers to vacuum up personal details about hundreds of millions of Americans. The companies are tracking your race, education level, TV habits, marital status, net worth, etc.. They're collecting what you post on social media, whether you're behind on your bills, what you order online. Then they feed this information into complicated computer algorithms that spit out predictions about how much your health care could cost them.

Find out more about this here.

Copyright © Designed & Developed by PrivacyEnhACT 2021

Figure 5.3: PrivacyEnhAction: Fitbit inferences from heart rate data

Alexia Dini

5.5 Discussion

In this Chapter, we have investigated the possibility of getting insights and extracting inferences about the users from their data collected from fitness trackers, aiming to provide an answer to Research Question **RQ2:** ***"What inferences can be made from data collected from smart home devices and fitness trackers?"***. In order to address this question, the literature review we performed in the area assisted us in the formulation of a list of possible inferences that pose a threat to user privacy when using fitness trackers. We limited the inference list to those inferences that we could identify at the time that the research was performed based on the available data available. Using the list, we implemented the new functionalities for the PrivacyEnhAction application for the three fitness trackers we had at our disposal, and the results showed that multiple data points can be used to infer and possibly predict health and fitness status, pregnancy, religion, etc. Not surprisingly, Prince in her work [229]) explains very effectively that a big amount of health information can be inferred from location data.

The methodology presented in this Chapter can be adapted in other scenarios as well as it is not bound to smart home or fitness trackers scenarios. We believe that the results of our

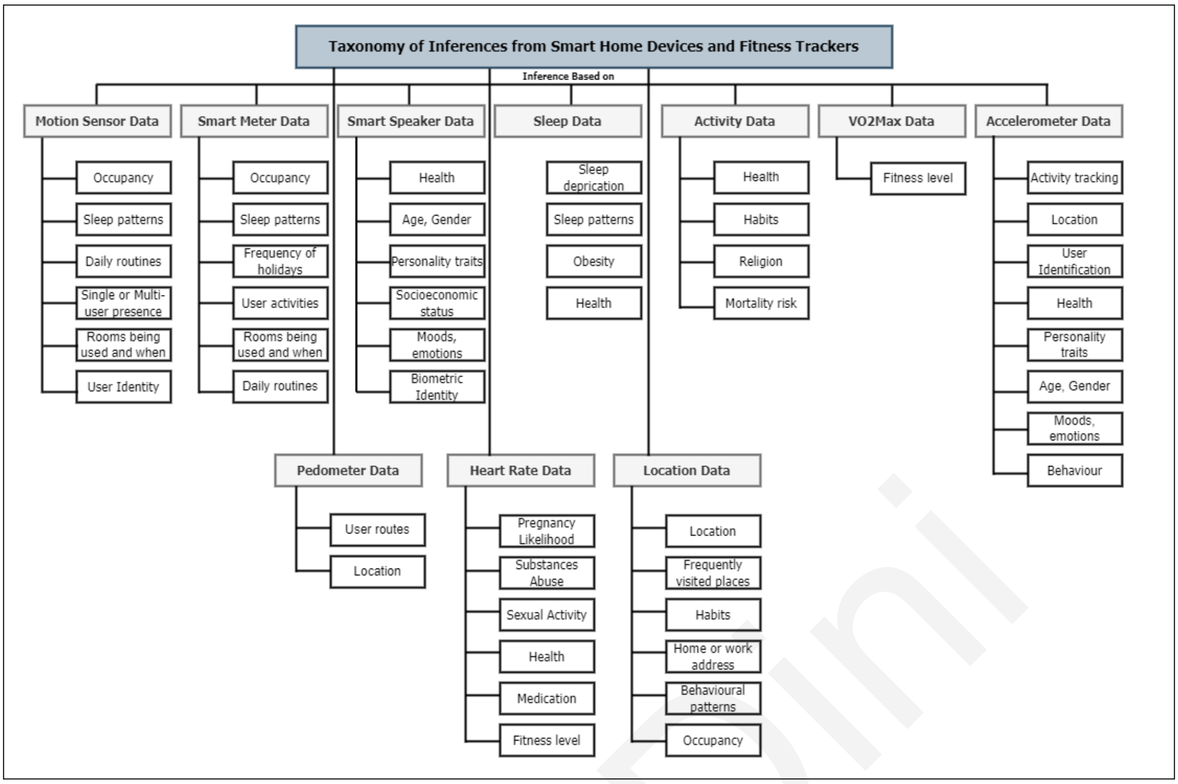


Figure 5.4: Taxonomy of Inferences from Smart Home Devices and Fitness Trackers

experimental research could can act as a stepping stone in a common effort to bring the smart devices owners in the heart of the privacy risks awareness process with the aim to increase their knowledge and guide their attention towards those actions that can protect them from potential harm, and also for the provision of better services to the users.

Based on the results from the literature review carried out for Chapters 4 and 5 and the research we performed towards answering RQ2, we have created a taxonomy for the inferences from smart home devices and fitness trackers, that can be seen in Figure 5.4, which is one of the contributions of this doctoral thesis. The taxonomy assembles together a summary of the information for the inferences gathered from the literature review and from the research conducted in this thesis, for presentation purposes.

Alexia Dini

Chapter 6

Evaluating the impact of PrivacyEnhAction to users awareness

Alexia Dini

The purpose of this Chapter is to present the findings of a survey that was conducted to evaluate the impact of the PrivacyEnhAction web application on users' awareness in relation to the privacy risks and the inferences that could be drawn about them from their fitness trackers data, as through our research in this thesis we have identified that there is an imperative need for the development of tools and user awareness mechanisms that can assist the users in understanding how the data created by their smart devices can be exploited for the extraction of inferences regarding their daily activities and lifestyle in general. The Chapter begins by introducing the research questions we aim to answer. It will then provide an overview of the methodology used to design and administer the two questionnaires involved and the approach used for performing the user evaluation. Next, the chapter will present and analyse the results of the questionnaires that were collected from the same users before and after interacting with the PrivacyEnhAction web application. Finally, the chapter will discuss the findings in relation to the research questions that were defined in this Chapter. The work presented in this chapter is based on our research published in the Springer *User Modeling and User-Adapted Interaction* journal [83].

6.1 Research Questions

One of the contributions of this thesis is PrivacyEnhAction, a web application that aims to provide awareness to the users about the possible privacy risks and the inferences that can be extracted about them from their fitness trackers and smart home devices data, so that they can set their user privacy preferences in such a way that their personal privacy can be protected. In our effort to address this goal, we define the following research questions.

RQ3: ***"Are the users aware of the inferences that can be made about them from their fitness trackers data?"***

For providing an answer to this research question, we conducted an online questionnaire that targets fitness trackers users in order to gain an understanding of: (i) their concerns over their privacy when using their devices, (ii) their awareness of what data are collected by their fitness trackers and how these are being used and shared, (iii) their awareness on the privacy risks from fitness trackers data.

RQ4: ***"Can we enhance the awareness of the users regarding the possible inferences that can be obtained from their fitness tracker data?"***

To answer this question we provided the same group of fitness trackers users with a number of datasets from three fitness trackers brands (Fitbit, Garmin and Xiaomi). The users were asked to use one dataset for each fitness tracker brand in order to interact with the PrivacyEnhAction web application and review the analysis results. Following that, they were required to complete an evaluation questionnaire about the app, where they were also expected to answer similar questions to the questionnaire used in RQ3, in order to gain an understanding of whether their awareness regarding inferences has been increased.

6.2 Material and methods: Empirical approach

In order to perform a user evaluation to assess the impact that the PrivacyEnhAction application can have to the awareness of the users, we followed a three-step empirical approach, that is described below.

1. **Step 1**: A first questionnaire whose aim is to collect information about the awareness and the concerns of fitness trackers users regarding their privacy when using fitness trackers was created and distributed (*Questionnaire on fitness trackers user privacy concerns* [159]).
2. **Step 2**: The participants were provided with the datasets collected during the data collection process described in Chapter 5, after they were anonymised, and were requested to use them in order to interact with the PrivacyEnhAction application. The existing datasets were used for evaluation purposes and in order to let participants use the application without providing their own personal data.
3. **Step 3**: A second questionnaire was created and distributed to the same group of users as in Step 1 and Step 2. By using a number of questions similar to the ones in the first questionnaire, the aim of this questionnaire was to assess if the users' awareness

Alexia Dini

and privacy related concerns have changed (i.e. improved) after interacting with the application (*PrivacyEnhAction Evaluation Questionnaire* [158]).

It has to be noted that the participants had to complete all the steps in order for their response to be considered as valid. For the analysis of the results we used IBM SPSS Statistics for the generation of data descriptive statistics and item-level results of each question.

6.2.1 Research participants recruiting

The User Evaluation survey was distributed through email communication in order to recruit participants. No monetary or other incentive was provided as a reward for answering the survey. The email provided information about the research goals, stating the objectives of the study and it also included the links to the survey questionnaires, the PrivacyEnhAction application and the share link of the available datasets and the application user guide. No screening criteria were applied, other than that the participants had to be owners of fitness trackers or smart watches. A total of 47 responses were collected. Out of these responses, 17 participants did not complete the second questionnaire and as such these data were removed. Finally, we ended up with a total of 30 valid responses which were used in the analysis.

Alexia Dini

6.3 Questionnaire on fitness trackers user privacy concerns

In the initial questionnaire, the first section consists of social and demographic questions, like gender, age, education level and profession. We used the gender as a demographic variable in order to determine if there exist any opposing views in the attitude and awareness of the privacy risks of the use of fitness trackers between male and female users of the study. In the literature, age is considered as a negative factor in the acceptance of technology [223], and for that reason we also used this as a demographic variable in order to find out if it can affect the results in relation to the awareness of the users of the privacy risks of the use of fitness trackers. The second section includes questions regarding information about fitness tracker ownership, such as frequency of using a fitness tracker, length of time of ownership of a fitness tracker and the fitness tracker brand being used. The third part consists of questions related to the user's attitudes towards reading the fitness tracker's privacy policies and changing the default privacy settings. The fourth section includes questions about the user awareness on fitness tracker data collection and sharing, while the fifth section consists of questions related to the user's awareness on the privacy risks form fitness trackers data. The

Table 6.1: Information about fitness tracker brands being used by the survey participants

Brand	**Frequency**	**Percent**
Apple	2	6.7
Fitbit	4	13.3
Garmin	6	20
Samsung	5	16.7
Xiaomi	5	16.7
Other	8	26.7
Total	**30**	**100%**

sixth part of the survey includes questions about the user's privacy concerns when using fitness trackers. The next section contains questions regarding the users' attitudes in relation to good uses of data if shared, and the last section gathers the user opinions about the importance of the creation of tools that would make the users aware of how their data are collected and shared by smart devices.

6.3.1 Demographics and other results

In the data analysis, the gender breakdown achieved was 66.7% male and 33.3% female. The majority of the participants are employed at the engineering and manufacturing sector (30%),the IT sector (26.7%), followed by the education (10%), accountancy, banking and finance (6.7%), business, consulting and management (6.7%), environment and agriculture (3.3%), healthcare (3.3%), and other sectors (13.6%). In relevance to the fitness tracker or smart watch brand being used, Table 6.1 shows the frequency and percentage of participants using each fitness tracker brand mentioned. The length of time that the participants have been using their fitness trackers or smart watches is reported in Table 6.2.

The analysis of the responses in the third section of the questionnaire that is related to the users attitudes towards reading the fitness trackers privacy policies and changing the default privacy settings, showed that 80% of the participants do not read the privacy policy of their fitness tracker, 86.6% do not read the terms and conditions and 70% have never changed the default privacy settings, as seen in Figure 6.1.

The aim of the next section of the questionnaire was to examine the participants' per-

Alexia Dini

Table 6.2: Information about length of time of owning a fitness tracker

Answer	**Frequency**	**Percent**
The past 3 months	6	20
The past 6 months	2	6.7
The past year	4	13.3
The last 2 years	4	13.3
More than 2 years	14	46.7
Total	**30**	**100%**

ceived knowledge and awareness of the data collection process performed by fitness trackers or smart watches, as well as to see if they acknowledge the types of data collected and also what happens to that data afterwards, using a 'Yes', 'No' and 'Maybe/I don't know' type of question. The results presented in Figure 6.2 show that a big percentage of the participants (83.3%) is aware that personal data are collected by fitness trackers, but only a 3.3% understands how these data are being used by the service provider and a 30% of the participants is aware of the types of data that are being collected by fitness trackers.

6.3.2 User awareness on privacy risks

In this section, in order to understand the users' awareness and perception of the possible privacy risks emerging from the use of fitness trackers, the participants were presented with a number of events and were asked to give their opinion regarding the possibility that they could occur, using a 5-point Likert scale with values ranging from 1 = "*Very unlikely to happen*" to 5 = "*Very likely to happen*".

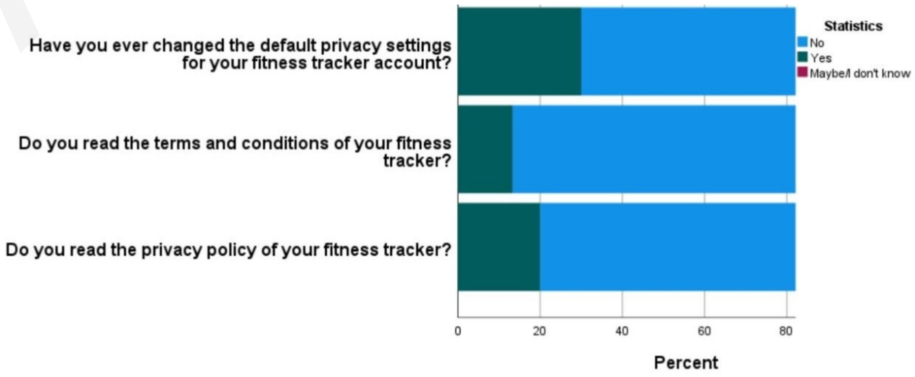


Figure 6.1: Users attitudes with regards to privacy policies

Alexia Dini

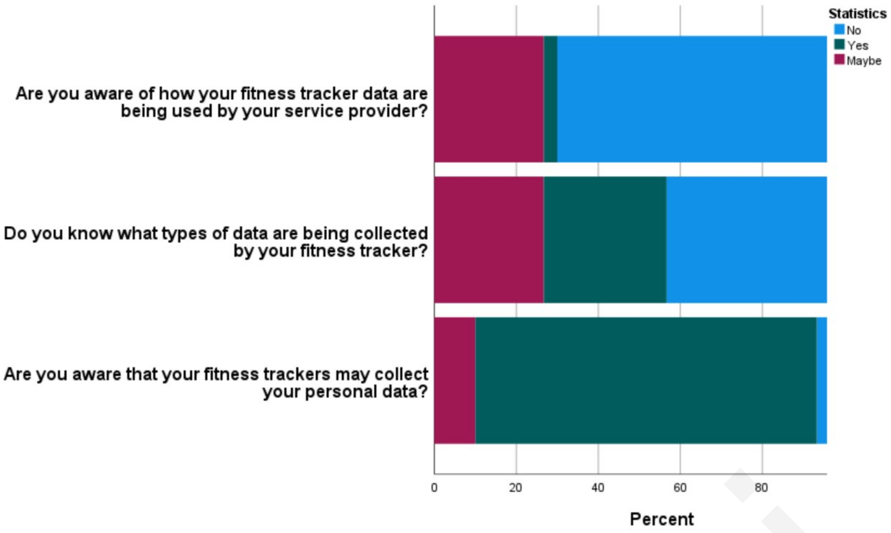
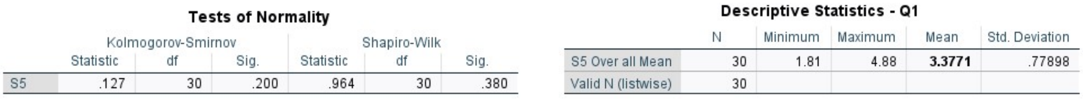


Figure 6.2: User awareness on fitness tracker data collection and sharing

Alexia Dini

The test of normality showed that the data is normally distributed with p=0.38. The overall mean score of the Likert scale in the fifth section of the questionnaire consisting of 16 items is 3.37. This score can be interpreted as the average response of the users in relation to their awareness about the possible inferences (in relevance to the scenarios they were presented with) to be “*Undecided*”. The normality test results and descriptive statistics of the data can be seen in Figure 6.3, while the actual events that were presented to the users, along with the users’ responses can be found in Figure 6.4. From these results, it appears that the participants are aware of a small number of inferences that can be drawn from their fitness tracker data. For example, in regards to the scenario *“Marketing companies can use fitness tracker data in order to send you specific advertisements regarding running shoes”*, 68.2% of the users reported this as “*Very likely to happen*" and 18.2% as “*Likely to happen*", while none of the respondents responded with “*Very unlikely to happen*" or “*Unlikely to happen*". This is quite predicted as online targeted advertising has shown great market potential [322] and is widely used today. In another case, the scenario *“A murder can be solved by using the victims fitness tracker data, such as heart rate data”* has been acknowledged as “*Very likely to happen*" by 54.5% and as “*Likely to Happen*’ by 27.3% by the participants (none of the participants responded with “*Very unlikely to happen*" or “*Unlikely to happen*"). This is explainable as in the recent past there have been many murder cases reported in the news where the data from the fitness tracker worn by the victim has assisted in the determination of the exact time of death and led to the murder being solved [124], [183].

The participants opinions diverged regarding religion inferences, as 22.7% have responded to this scenario as “*Very unlikely to happen*" and 13.6% as “*Unlikely to happen*", but a 40.9%



Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
S5	.127	30	.200	.964	30	.380

Descriptive Statistics - Q1

	N	Minimum	Maximum	Mean	Std. Deviation
S5 Over all Mean	30	1.81	4.88	**3.3771**	.77898
Valid N (listwise)	30				

Figure 6.3: Normality test and descriptive statistics on Questionnaire 1 Section 5 data

is undecided about this possibility. Similar levels of responses across all answers were observed for the scenario *"Your fitness tracker data can be used to make the assumption that you are an alcoholic"*, where the answers were spread with 22.7% for "*Very likely to happen*" and "*Likely to Happen*", 18.2% for "*Undecided*", 13.6% for "*Unlikely to happen*" and 18.2% for "*Very unlikely to happen*".

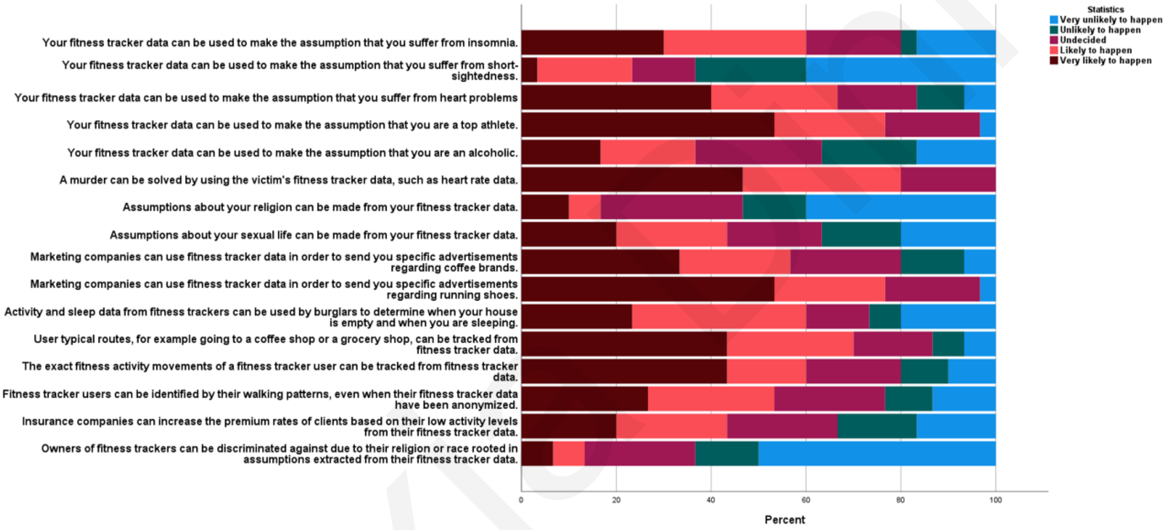


Figure 6.4: Users responses on Questionnaire 1 Section 5

In relation to the effect that the participants' gender has to answers, further analysis on this section's questions has shown that the gender is not correlated with the user awareness about the possible inferences that can be extracted from fitness trackers data. Furthermore, using the ANOVA test, we investigated the effect that age has on the responses, and we deduced that age has a significant impact on the following statements:

1. *Insurance companies can increase the premium rates of clients based on their low activity levels from their fitness tracker data* (F=3.335, p=0.026): For this scenario, younger participants (aged 18-25) have responded with a mean score of 4.67, thus showing that they believe that such a scenario is very likely to happen, while older participants (aged 56-65) have responded to this question with a score of 1, i.e. as very unlikely to happen and participants aged between 46-55 have responded with unlikely to happen.
2. *Marketing companies can use fitness tracker data in order to send you specific ad-*

vertisements regarding running shoes (F=5.477, p=0.003): In this scenario younger participants believe that it is very likely to happen, while older participants are more reluctant to accept it.

3. *Marketing companies can use fitness tracker data in order to send you specific advertisements regarding coffee brands* (F=2.941, p=0.04): Older participants believe that this scenario is likely to happen while younger participants are more sceptical.

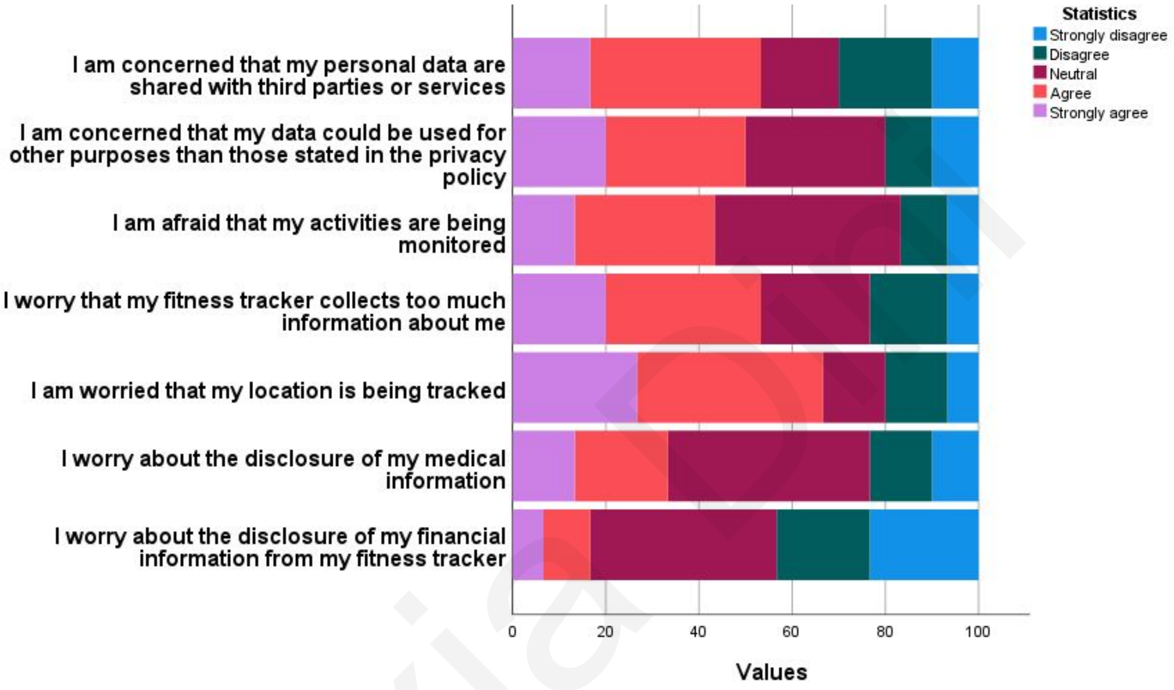


Figure 6.5: Examples of user privacy concerns regarding the use of fitness trackers

6.3.3 User privacy concerns

In order to understand the privacy concerns of the participants, they were asked a number of questions about specific concerns related to the use of fitness trackers, using a 5-point Likert scale with values ranging from 1 = "*Strongly disagree*" to 5 = "*Strongly agree*". The concern that worries the participants the most is the possibility that their personal information may be used for targeted advertising, where 33.3% of the participants strongly agree and 23.3% agree with the statement, followed by the fear that their location in being tracked, with 26.7% of the participants responding with Strongly agree and 40% with Agree (Figure 6.5). Further analysis on the questions in this section shows that gender does have an effect on the users' privacy concern.

In relation to the participants' awareness to the data collected by fitness trackers, location is the most popular answer to this open question (60%), followed by heart rate (40%) activity

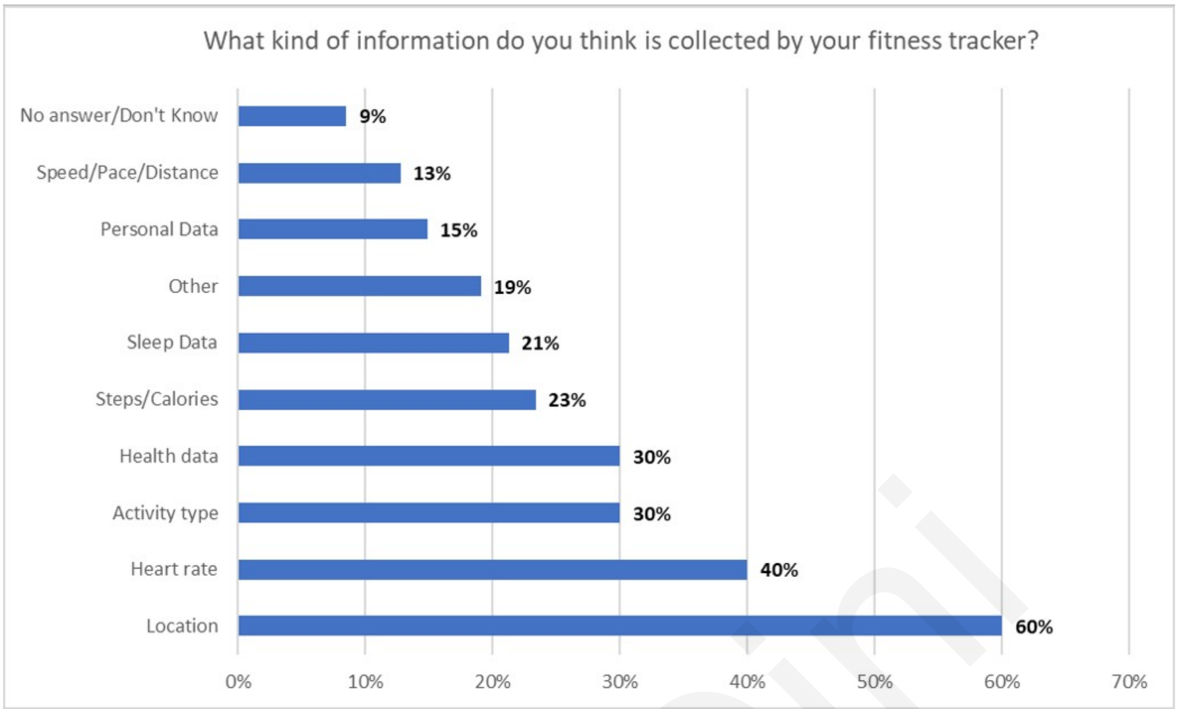


Figure 6.6: User awareness regarding the data collected by fitness trackers

type (30%) and health data (30%) (Figure 6.6).

6.4 PrivacyEnhAction Evaluation Questionnaire

6.4.1 PrivacyEnhAction usability results

The first part of the second questionnaire aims to measure the usability of the PrivacyEnhAction application using the System Usability Scale, which is a stable, efficient and valid way to calculate the usability of a system [43, 293]. The SUS score for PrivacyEnhAction was calculated using the participants' responses to be 83.75, which is considered to be an excellent SUS score, as it is well above the average SUS score, which is 68.0 [43].

6.4.2 User awareness on privacy risks

In the second questionnaire, the participants had to answer the same set of questions regarding their awareness on the privacy risks and the possible inferences that could be extracted from fitness trackers data, as in the first questionnaire, after their interaction with the PrivacyEnhAction application, in our effort to seek an answer to Research Question **RQ4:** ***"Can we enhance the awareness of the users regarding the possible inferences that can be obtained from their fitness tracker data?"*** In Section 6.3.2, we showed that the overall mean

Alexia Dini

Table 6.3: Overall mean scores on user awareness before and after interacting with the PrivacyEnhAction application

	Min	**Max**	**Mean**	**SD**
Before	1.81	4.88	**3.33771**	.77898
After	3.00	5.00	**4.0417**	.50704

score of the same section in the first questionnaire regarding user awareness on privacy risks was 3.38. The overall mean score of the same set of questions in the second questionnaire, i.e. after the participants have interacted with the PrivacyEnhAction application, is 4.04, as can be seen in Table 6.3.

To verify our results, we conducted a paired sample t-test in order to compare the degree of the users' privacy awareness before and after interacting with the PrivacyEnhAction application, using the same set of questions that exist in both questionnaires regarding user awareness on privacy risks. Using Cronbach's alpha indicator, we evaluated the reliability of the two Likert scale sets of questions of the questionnaires. The results for the first questionnaire demonstrated good internal consistency with a score of 0.876, while the results for the set of questions of the second questionnaire showed acceptable internal consistency with a score of 0.768. The results of the paired sample t-test suggest that there is a statistically significant difference between the level of the users' awareness before and after their interaction with the PrivacyEnhAction application, as shown in Table 6.4. A p value below 0.05 was considered statistically significant.

The pairs of questions that differ before and after the users' interaction with the PrivacyEnhAction application are the following:

- **Pair 1:** *Owners of fitness trackers can be discriminated against due to their religion or race rooted in assumptions extracted from their fitness tracker data.*
- **Pair 2:** *Insurance companies can increase the premium rates of clients based on their low activity levels from their fitness tracker data.*
- **Pair 4:** *The exact fitness activity movements of a fitness tracker user can be tracked from fitness tracker data.*
- **Pair 6:** *Marketing companies can use fitness tracker data in order to send you specific advertisements regarding running shoes.*
- **Pair 9:** *Assumptions about your religion can be made from your fitness tracker data.*

Alexia Dini

Table 6.4: Paired-samples t-test results

Question Pair	Mean	SEM	t	df	Sig
Pair 1	-1.7	0.3	-5.667	29	**0.001**
Pair 2	-0.967	0.309	-3.13	29	**0.002**
Pair 3	-0.5	0.302	-1.654	29	0.054
Pair 4	-0.767	0.261	-2.935	29	**0.003**
Pair 5	-0.433	0.27	-1.606	29	0.06
Pair 6	-0.433	0.223	-1.941	29	**0.031**
Pair 7	-0.167	0.369	-0.452	29	0.327
Pair 8	-0.433	0.345	-1.257	29	0.109
Pair 9	-1.733	0.332	-5.222	29	**0.001**
Pair 10	0.133	0.243	0.548	29	0.294
Pair 11	-0.867	0.321	-2.703	29	**0.006**
Pair 12	0	0.275	0	29	0.5
Pair 13	-2	0.303	-6.595	29	**0.001**
Pair 14	-0.633	0.305	-2.076	29	**0.023**
Pair 15	-2.267	0.325	-6.975	29	**0.001**

- **Pair 11:** *Your fitness tracker data can be used to make the assumption that you are an alcoholic.*
- **Pair 13:** *Your fitness tracker data can be used to make the assumption that you suffer from short-sightedness.*
- **Pair 14:** *Your fitness tracker data can be used to make the assumption that you suffer from heart problems.*
- **Pair 15:** *Your fitness tracker data can be used to make the assumption that you suffer from insomnia.*

We further analyse if the users will take specific actions after their interaction with the application in relation with the use of their fitness trackers. In particular, 53.3% of the participants said that it is *very likely* that they will change the default privacy settings of their tracker, while 23.3% responded that this is *likely*. Regarding the statement *"Allow the tracker provider to use your data for specific purposes that you choose"*, 56.7% and 23.3% of the participants responded that this is *very likely* and *likely* to happen, respectively. If we compare the participants answers in percentages in Figure 6.7 with their responses in Section 6.3.3 in relation to the participants' attitudes against fitness trackers privacy policies, terms and conditions, etc., we can see that PrivacyEnhAction has increased their awareness, as 26.6% more of the participants will now read the privacy policy of the trackers, 36.7% more will now read the terms and conditions and 46.6 % more will now change the default privacy settings of their account.

Alexia Dini

6.4.3 User opinions about PrivacyEnhAction

In the next section, the participants had to provide their feedback with regards to their interaction with the PrivacyEnhAction application. According to the responses, 83.3% of the participants think that their awareness regarding the use of their personal data from their fitness trackers has increased after they have used the app. Furthermore, 56.7% of the respondents find that their awareness about the possible inferences that can be made about them and their habits from their fitness trackers data has increased to a high degree, while 30% think that it has increased very much.

Regarding the users' privacy concerns, 86.6 % of the participants think that the use of the PrivacyEnhAction application has increased their awareness about the use of their personal data ranging from *very* to *a high degree*, while 10% think that it has not increased their awareness at all. It is however very important to mention that all the participants have reported that they believe that PrivacyEnhAction is a useful tool for informing them about the possible inferences that can be extracted about them from their data that may violate their privacy and to provide user awareness.

Alexia Dini

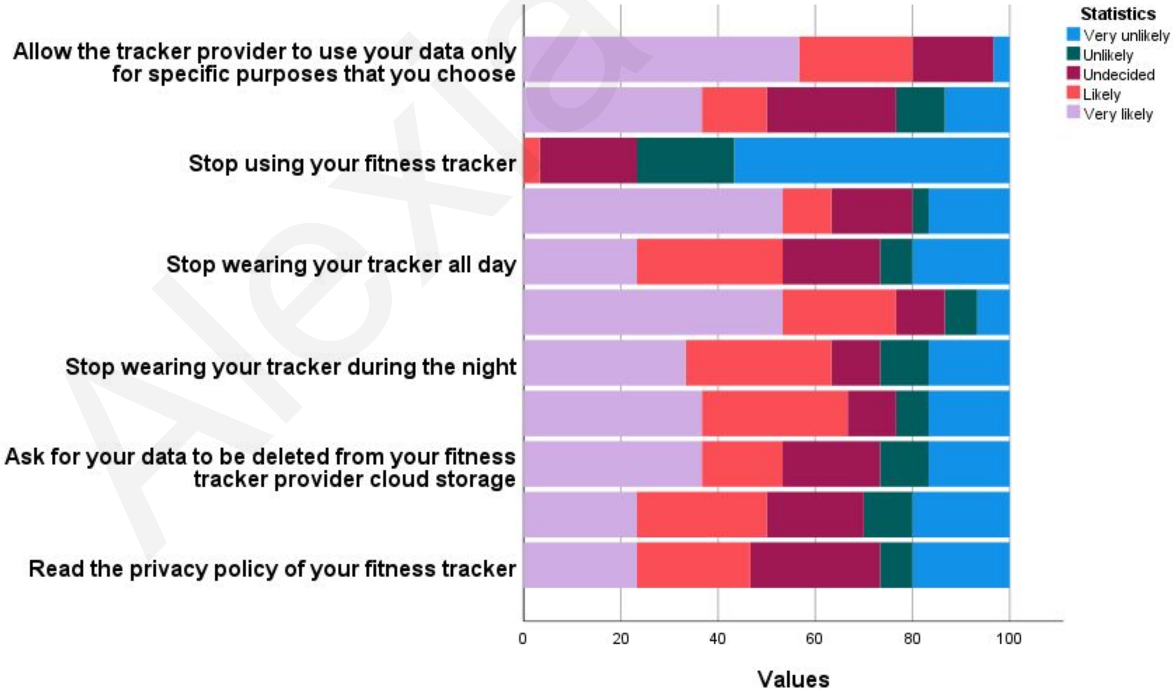


Figure 6.7: Examples of users possible actions after using PrivacyEnhAction

6.5 Discussion

In Chapters 5 and 6, our research was guided by the ambition to create a tool that will increase the users' awareness in the area of fitness trackers with reference to what information can be extracted about them from the data collected and shared by their fitness trackers. Our intention was to educate the users about the possible risks and enable them to set their privacy preferences on their fitness trackers accordingly.

The literature review and the research we performed at this stage assisted us in the formulation of a list of possible inferences that pose a threat to user privacy when using fitness trackers, which we used to implement new functionalities in the PrivacyEnhAction application for the three fitness trackers under study. The findings of the study presented in this Chapter demonstrate that the use of tools, like PrivacyEnhAction, can assist in the enhancement of the users' privacy awareness when using smart devices.

In our work, we aimed to gain an understanding of fitness trackers users' awareness and concerns regarding their privacy when using fitness trackers, through the first questionnaire. The results have shown that even though a big percentage of the users are aware that their trackers may collect their personal data, they are not aware of the inferences that can be extracted about them from their fitness trackers data and as such, they do not take any action to minimise any possible risks, as for example by altering their fitness trackers privacy preferences or by reading the privacy policies of their trackers in order to get informed. This finding agrees with prior research in the area, where results show that fitness tracker users do not change the default settings of their devices and they do not read their privacy policies [302], [109], even though the majority of the respondents agrees with the privacy policies and terms of service, they continue to skip them due to information overload [260], and also because they consider them to be annoying and lengthy [211]. This observation also indicates that personal data privacy awareness is not equivalent to the understanding of personal data privacy protection [55]. We also found that only a small portion of the sample understands how the personal data collected by fitness trackers are being used by the service providers. This is in line with the work of Vitak et al. [304], which showed similar results from a survey of Fitbit and Jawbone users about the user privacy concerns in relation to tracking and sharing.

Our participants responses in relation to their awareness about the inferences that could be extracted from their data and how these could be used by third parties, showed that the users are apprehensive only for a few of the scenarios that they were presented with, while

Alexia Dini

overall they seem uncertain about the possibility of the extraction of the presented inferences. A previous study in the area by Velykoivanenko et al. [302] has linked the participants beliefs with their understanding of the embedded sensors in their device and the data collected by those sensors. This could justify the participants responses in relation to the scenarios presented to them, and enable us to give an answer to Research Question **RQ3:** ***"Are the users aware of the inferences that can be made about them from their fitness trackers data?"***, where we can say that the user awareness depends on the scenario, but in general the users are not aware of the possible inferences that could be extracted about them.

The results of the analysis of the second questionnaire have produced more comprehensive conclusions as to whether the users' interaction with the necessary tools can increase their awareness regarding the possible inferences that can be obtained from their fitness tracker data (**RQ4**). In regards to the inferences that could be extracted from the users' data or how these data could be used by third parties, it has been observed that the participants seem to be more educated and more aware about them after interacting with the application, as the mean value of the responses in the relevant section of the questionnaire is "*Likely to happen*". Comparing this with the mean value "*Undecided*" in the same set of questions from the first questionnaire, we can conclude that the users' interaction with the PrivacyEnhAction application has increased their awareness regarding the inferences that could be extracted from their data and how these data could be used by third parties. This demonstrates that the privacy education that PrivacyEnhAction intents to bring to the users through its graphical interfaces, the pop up messages and the educational tips it provides, seems to be working, and proves that embedding privacy education in an application with simple and clear descriptions is a required feature for enhancing user privacy awareness and education [302], [5]. As users appear to be ignorant of how their personal data could potentially be used, it is important that education mechanisms take the context into consideration when including the user in the process. For the fitness trackers examples under study, this is essential due to the sensitive types of data collected.

Our results showed a positive relationship between the use of a privacy awareness mechanism and the increase of the awareness of the user about the possible privacy risks of using a fitness tracker. Enhancing the users' control over their privacy by assisting them to understand the data practices of the smart devices they own, adds to the strengthening of their privacy awareness. These findings are aligned with earlier studies where it is reported that privacy awareness mechanisms like data dashboards, similar to PrivacyEnhaction, are well perceived by users in terms of effectiveness and easiness to use, and also due to the detailed

Alexia Dini

information provided [282]. The communication of the potential privacy risks to the users and its effect to the users' awareness is also investigated in our study. The results showed that the users' privacy awareness had a positive relationship with informing the users about any potential privacy risks, being in line with previous studies which give directions for the creation of privacy awareness mechanisms [303].

The findings in this Chapter of the thesis provide valuable insights for the users of fitness trackers in our effort to increase their awareness, however despite the possible privacy risks, the inferences that can be extracted from fitness trackers data can also have a positive impact to the users. Tracking the daily activities of a user can help to enhance the user's health in the long term as the user can be assisted to reach her fitness goals [320]. The observation of personal health data collected from fitness trackers can lead to the detection and prevention of diseases, such as Covid-19 [117], heart diseases [144], [7], or diabetes [21], and even sleep problems [248]. In all cases, it is important that the users understand the privacy complications of using fitness trackers and the potential inferences from personal data, while at the same time balancing the benefits of their functionalities.

Limitations. We acknowledge that the survey presented in this Chapter may have some limitations, however it could provide the means for further research in the relevant area. First, the size of the participants sample cannot represent the smart devices user population, even though we tried to recruit a diverse sample of participants in terms of demographic variables in order to increase the probability that the results we are aiming for have been indicated by at least one of our participants. Hence, the statistical analysis performed on our sample provides only indications; it is, however, useful in analysing our results. Even though our participant recruitment methods were designed to minimise response bias, by electronic mails to random and known addresses at public and private universities at Cyprus and abroad, the sample is considerably more educated than the general population. This parameter may bring bias to the results in terms of the knowledge and the awareness of the users regarding the privacy risks.

Another limitation is the reluctance of a portion of users to be educated about the privacy risks of using a fitness tracker, as they consider that the benefits of their devices are more important than any possible risks, and are therefore uninterested in anything other than the provided services. When starting our research, we acknowledged that these types of users will probably not going to use the PrivacyEnhaction application. In order for the users to seek technologies or applications that educate them about fitness trackers privacy risks, policy makers and regulatory organisations should engage in actions aiming to increase the privacy

Alexia Dini

awareness of users of smart devices in general. To that end, it is essential to provide tools and methods that enable the increase of privacy awareness.

Alexia Dini

Chapter 7

Analysis of smart devices privacy policies

Alexia Dini

Regulations, such as the GDPR, oblige service providers to inform the users about their practices regarding data collection and processing [280] and the existing method used for the portrayal of the rights and responsibilities of both the user and the service provider in terms of data collection, processing and sharing, are the privacy policies, which are legal texts that depict the practices that an organization or company follows when handling the personal data of its users [233]. As transparency and user consent are essential factors in GDPR, Article 12 states that any communication relating to data processing must be provided to the user *"in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child"*.

In Chapter 6, the results of the survey conducted showed that 80% of the participants does not read the privacy policy of their fitness tracker. Based on this, we decided to take a step back in order to investigate what the privacy policies of smart home devices and fitness trackers mention in their text regarding the data collection, processing and sharing practices. In this Chapter, we aim to add to our effort to further increase the awareness of fitness trackers users and particularly smart home devices users, something that has not been examined in Chapter 6, by examining the privacy policies of the three fitness tracker brands used in the study in Chapters 5 and 6, and those of a number of smart home devices. We will look into what data these devices collect, how these data are used, and who can access them. By understanding the privacy policies of these devices, users can make informed decisions about which devices to use and what data to share. This analysis aims to provide clarity and transparency about the privacy practices of these devices, ultimately empowering users to make informed decisions about their data and privacy.

7.1 Fitness trackers privacy policies

Fitness trackers assist the users in tracking their health, by enabling them to specify what they want to record about themselves, such as their weight, the exercise they perform, the number of steps they take during the day, the distance they walk, how much and when they sleep, their heart rate, etc. This stored information is clear to the users, as these are the data they can see through their profile dashboard. However, further user information is accumulated from the trackers that the users may be unaware of, like their location, timezone, IP address, etc. Even though fitness trackers privacy policies usually state that no data are shared with third parties, this is not always the case as constant user tracking and data collection give fitness tracker companies the opportunity to capitalise on user data with the help of third party sales [49].

But what do the privacy policies of the fitness trackers used in this study state regarding data sharing? In this section, we provide a review of how Fitbit, Garmin and Xiaomi fitness trackers address data sharing in their privacy policy.

For the review, we have used as an example the work performed by Perez et al. [226]) where the authors have performed an analysis of the privacy practices that manufacturers provide related to data collection, data ownership, data modification, data security, external data sharing, policy change and policies for specific audiences for six IoT devices and systems, including Fitbit devices. Based on this analysis, we have followed a methodology for gathering the required information about data collection, data sharing, data recipients, privacy policy changes and data handling in case of reorganisation/merge/resale, extending the review to the areas of our research interest. A summary of the privacy policies review can be seen in Table 7.1.

Alexia Dini

7.1.1 FitBit privacy policy regarding data sharing

The Fitbit Privacy Policy states that: *"We never sell the personal information of our users. We do not share your personal information except in the limited circumstances described below."*[1]. The listed circumstances are: (i)when the user agrees to use FitBit community features like forums, challenges, or social tools, or directs FitBit to share her data with third parties, as for example when the user gives a third-party application access to her account, or provides access to her employer when choosing to participate in an employee wellness

[1] https://www.fitbit.com/global/us/legal/privacy-policy

Table 7.1: *Comparison of FitBit, Garmin and Xiaomi fitness trackers privacy policies*

FT	Data collection	Data sharing	Data recipients	Privacy policy change	Data handling in reorg.
Fitbit	Account, location, usage, biometrics and fitness data (steps, distance, calories, weight, heart rate,sleep stages, active minutes).	Only shares personal data when user agrees to share, for external processing with their partners according to their policies, or for legal reasons and to prevent harm.	3rd party apps, FitBit corporate affiliates, service providers, and other partners.	Fitbit will notify the users before making any changes to the privacy policy and the users will be able to review the revised policy before deciding if they would like to continue to use the Services.	Users are informed that adequate measures will be taken to protect the confidentiality of personal information. Affected users will be given notice before transferring data to new entity.
Garmin	Account, health, fitness data (step count, heart rate, sleep data), activity data (runs, bike rides, swims, or other activities recorded with the device).	Only shares data with 3rd party apps, platforms or service providers with whom the users ask Garmin to share their data with.	Third party apps, platforms, service providers.	Garmin may update their Privacy Policy. The users will be provided with notice when this is required by applicable law and will be asked to provide their consent.	Users personal data may be transferred to a new entity provided that this will not be permitted to process personal data as per the privacy policy, without first providing notice to users and obtaining consent.
Xiaomi	Activity data, sleep, blood oxygen saturation information, heart rate, weight, call records, telephone number, SMS content, contact name and caller number, MAC address, serial number, firmware version, system time, mobile phone operating system version, brand model, information submitted via services, information about NFC, other information.	May disclose user personal information to 3rd parties to provide requested products/services. May also disclose personal information to other affiliated companies, to comply with legal obligations, to protect and defend their rights and property or with the user permission.	Xiaomis ecosystem companies, 3rd party service providers and business partners, other 3rd parties.	If there are material changes to the privacy policy, Xiaomi will notify the users by email, or post the changes on their websites or through their software.	Users information may be sold or transferred as permitted by law and/or contract. The users will be notified via email and/or a prominent notice on Xiaomi's website of any changes in ownership, uses of their personal information, and choices they may have regarding their personal information.

Alexia Dini

program, (ii) for external processing, to their partners who process user data on FitBit's behalf in compliance with its policies, and (iii) for legal reasons or to prevent harm.

Even though Fitbit's privacy policy states that *"we never sell your personal data"*, it also states later that user data is used for marketing. What this means according to a Fitbit spokesperson is that user data is used only for advertising their own products [195]. In the case of a merger, acquisition, or sale of assets, the FitBit privacy policy informs the users that adequate measures will be taken to protect the confidentiality of personal information and give affected users notice before transferring any personal information to a new entity.

According to the Common Sense Privacy Program[2], a program that evaluates popular applications and services for children aiming to protect child and student privacy, Fitbit fitness trackers do not meet the organisation's recommendations for privacy and security practices. Some of the arguments behind this are, among others, that the trackers collect personally identifiable information (PII), that it is not clear if the data collection or use is bound to the requirements of the device, that the trackers collect geolocation and biometric or health data, and also that third parties collect user personal information.

The Garmin Privacy Policy includes in the list of possible recipients of the users' personal data various third party apps, platforms or service providers with whom the users ask Garmin to share their data. In these cases, the third partys handling of the users' personal data is the responsibility of that third party and the users are warned that they should carefully review the third partys privacy policy.

Additionally, Garmin's privacy policy states that: *"From time to time, we share or sell activity data in a de-identified and aggregated manner with or to companies that provide Garmin and our customers with content or features for the purpose of enhancing the quality of the content or features they provide and with or to other third parties for research or other purposes"*[3]. Regarding the possibility of any reorganisation, merger, or sale, the Garmin privacy policy clarifies that they may transfer users' personal data to an affiliate, a subsidiary, or a third party provided that any such entity will not be permitted to process personal data other than as described in the Privacy Policy without providing first notice to the users and obtaining their consent.

The Common Sense Privacy Program has only evaluated Garmin Vivofit Jr. and this specific device does not meet the organisations recommendations for privacy and security practices, for reasons such as the collection of PII, the possibility that user information can

[2] https://privacy.commonsense.org/

[3] https://www.garmin.com/en-US/privacy/global/

Alexia Dini

be transferred to a third party for advertising, marketing or other purposes, etc.

7.1.2 Xiaomi privacy policy regarding data sharing

The Mi Privacy Policy states that: *"We do not sell any personal information to third parties. We may sometimes share your personal information with third parties (as described below) in order to provide or improve our services, including offering services based on your requirements. If you no longer wish to allow us sharing this information, please contact us at https://privacy.mi.com/support"*[4]. The list of third parties includes Xiaomi's ecosystem companies, which are independent entities, other third party service providers and business partners who may have their own sub-processors, and other third parties with whom Xiaomi may share information in aggregated form. In particular: *"To help us provide you with services described in this Privacy Policy, we may, where necessary, share your personal information with our third party service providers and business partners. This includes our delivery service providers, data centres, data storage facilities, customer service providers and marketing service providers and other business partners. These third parties may process your personal information on Xiaomis behalf or for one or more of the purposes of this Privacy Policy....There may be occasions that third-party service providers have their sub-processors. To provide performance measurement, analysis, and other business services, we may also share information (non-personal information) with third parties in aggregated form"*. A worrying aspect of the privacy policy is that Xiaomi does not explain what the status of the users' personal information will be in the case of a merger, acquisition, or sale, as the only clarification given is that the users will be notified.

According to the Mozilla Foundation[5], Xiaomi's MiFit Smart Bands do not meet their Minimum Security Standards as they have not responded to how they handle security vulnerabilities. On top of that, Xiaomi has come under fire as it has been secretly collecting personal data from users of its products, and for these reasons the Mozilla Foundation warns the users against wearing these fitness bands[6].

Alexia Dini

[4] https://www.mi.com/uk/about/privacy/

[5] https://foundation.mozilla.org/en/

[6] https://foundation.mozilla.org/es/privacynotincluded/mi-band-5/

7.2 Smart home devices privacy policies

Smart home devices allow the users to control their environment, assisting in the enhancement of home automation and security. These devices continuously collect data such as location, videos, voice recordings, house maps and temperatures, movement patterns, electricity and water consumption, among others, while these data are analysed for the provision of services, with the aim to make the users lives easier. However, this analysis allows sensitive information to be inferred about the users, raising additional privacy concerns. Furthermore, the disclosure of smart home data data enables the profiling of users, as well as attackers or hackers or to perform targeted attacks. Smart home device manufacturers provide privacy policies with their products that aim to inform the users about the data collection, storage, and sharing practices of the provider, but in most cases the users tend to ignore those policies as they are lengthy.

In this section, we provide a review of the Vivint, Arlo, Hive and Philips Hue devices privacy policies, in order to understand how they address data collection and sharing. We have followed the same methodology we used as in 7.1 for gathering the required information about data collection, data sharing, data recipients, privacy policy changes and data handling in case of reorganisation/merge/resale. A summary of the smart home devices privacy policies reviewed can be seen in Table 7.2.

Alexia Dini

7.2.1 Vivint privacy policy regarding data sharing

Vivint smart home devices include smart home sensors, smart locks, smart thermostats and smart lights. The Vivint privacy policy states that *“We may disclose personal information described above with our affiliates and subsidiaries, service providers who act on our behalf, our business partners, third parties pursuant to legal purposes, or to others at your direction"*. The types of personal information described in the policy are very broad, including identifiers, such as name, physical address, email address, telephone number or account name, customer records, such as signature, physical characteristics or description, home or property information and credit-related information, and protected class and demographic information, such as age, sex, gender, disability, marital status, veteran status, race, and ethnicity. However, the privacy policy states that sensitive information, such as social security number, driving license number, passport number, debit or credit card number, in combination with any required security or access code, password, or credentials allowing access to

Table 7.2: *Comparison of Vivint, Arlo, Hive and Philips Hue privacy policies*

SHD	Data collection	Data sharing	Data recipients	Privacy policy change	Data handling in reorg.
Vivint	Identifiers, customer records, protected class and demographic information, device information, internet or other electronic network activity information, geolocation data, sensory information, professional or employment-related information, drawn inferences. sensitive information.	For business and marketing purposes, to provide and improve products and services, for risk management purposes, to comply with legal requirements or orders, to enforce their terms, conditions, and policies, or as per user consent.	Affiliates, subsidiaries, service providers, business partners, law enforcement or other government authorities.	No notification	Users are informed that their data will be disclosed as part of a corporate transaction.
Arlo	Contact details, login details, order and payment details, usage information.	To provide services, at the request of public authorities, in the context of court proceedings, mergers and acquisitions.	Companies belonging to the Arlo group, companies providing outsourced customer support, billing providers, cloud service providers, sales support providers.	Users will be notified of the changes within a reasonable time prior to the changes taking effect.	Users personal data will be shared with the acquirer.
Hive	Name, email address, telephone number, address, account details, transaction information, credit/debit card details, bank account details, device information.	Provision of services, data analytics and statistical research, development of new products and services, marketing .	Other companies apps, service providers, debt collection agencies, advertising parties, companies involved in providing the users with insurance, market research partners, government or regulators, police and law enforcement	If there are updates to the privacy policy, the version date will be updated.	The user personal data will be transferred to buyers who Hive sells or negotiates to sell all or part of their business or operations to.
Hue	Account, purchase, usage data, user generated content,third-party data.	For the provision of services, when required by law.	Affiliates, service providers, business partners, public and governmental authorities, professional advisors (banks, insurance companies, auditors, lawyers, accountants), other 3rd parties.	No user notification.	Users information will be shared with a business or another company involved with the sale of the business.

Alexia Dini

an account, precise geolocation, and racial or ethnic origin are also collected. Furthermore, the policy also states that: *“We may also combine the data we collect through Tracking Technologies with other data we collect to create a profile about you. We may disclose this profile data to business partners and ad networks so that they can show you advertisements that they think will interest you. Our partners may also use this information to recognise you across different channels and platforms, including but not limited to, computers, mobile devices, and smart TVs, over time for analytics, attribution, and reporting purposes"*. Regarding the possibility of any reorganisation, merger, or sale, the Vivint privacy policy does not include any information in the privacy policy with regards to what the status of the users personal information will be.

7.2.2 Arlo privacy policy regarding data sharing

Arlo’s smart home devices range include cameras, doorbells, floodlight cameras and various smart home accessories. Regarding data sharing, the privacy policy states that: *“We may share your personal data with reliable external parties, such as to other group companies, IT providers and companies with whom we partner to provide our services. We may also need to disclose personal data at the request of public authorities or to other parties in the context of court proceedings, mergers and acquisitions or similar"*. Examples of parties that user data can be shared with are companies belonging to the Arlo group, companies providing outsourced customer support, billing providers, cloud service providers, sales support providers etc. Arlo also claims that they will not sell the users’ personal data to any other party. However, they use profiling for data analysis and market research, which means that they process user personal data to evaluate personal aspects, get insights and make prediction about the users. In relevance to the possibility of any reorganisation, merger, or sale, the only information regarding what the status of the users’ personal information will be after the reorganisation provided by the privacy policy is that all the the categories of users personal data will be shared with the acquirer.

7.2.3 Hive privacy policy regarding data sharing

Hive is a company providing a range of smart home devices, such as smart thermostats, smart lights, motion sensors, door sensors and pet sensors. In their privacy policy, Hive states that regarding data sharing, user data are shared with other companies in the group of companies that Hive belongs to, while outside this group user data are shared with other

Alexia Dini

companies apps and products, installers and service engineers, various service providers, debt collection agencies, advertising parties, companies involved in providing the users with insurance, market research partners, government or regulators, police and law enforcement. Additionally, it is stated that the user personal data will be transferred to buyers or perspective buyers who Hive sells or negotiates to sell all or part of their business or operations to.

7.2.4 Philips Hue privacy policy regarding data sharing

Philips Hue provides a list of smart lights products and smart accessories. Regarding data sharing, the privacy policy of Philips Hue specifically states that *"We do not sell or rent your personal data. We share your personal data only when required to by law, if you provide us with permission, or to other parties acting on our behalf"*. According to the policy, user personal data are disclosed to affiliates, service providers, business partners, public and governmental authorities, professional advisors, such as banks, insurance companies, auditors, lawyers, or accountants, and other third-parties in the Philips Hue ecosystem. In the case of any reorganisation, merger, joint venture, or other disposition of the Philips Hue business, assets, or stock including in connection with any bankruptcy or similar proceeding, the user personal data will be shared with a business or another company involved with the sale of the business.

Alexia Dini

7.3 Discussion

In order to reinforce our effort to raise the awareness of the users of smart home devices and fitness trackers about the possible privacy risks of using these devices and the inferences that may be extracted about them from the data collected, we have performed a review of how a number of fitness tracker brands and smart home devices address data collection and sharing, and how these are presented in their privacy policies. Even though privacy policies should assist the users to make informed decisions regarding the use of their device, current policies lack usability, as users tend to ignore them and thus miss important information which includes details about providing their consent [241].

In regards to data collection, Fitbit and Garmin fitness trackers collect account, health, fitness, geolocation and device information, like number of steps, distance, calories, heart rate, weight, sleep stages, minutes active, as well as additional information that the users choose to provide. These types of information have been exploited in this thesis in order to increase the user awareness about the inferences that may be extracted about them from the data collected from their devices. In the case of the Xiaomi fitness trackers, the information collected far exceeds the necessary information for the service a fitness tracker is supposed to offer, as the devices also collect the MAC address, serial number, firmware version, system time and operating system version of the mobile phone connected with the Xiaomi Wear App, as well as information about SMS or message reminder functions, call records for making and receiving calls, the number of the mobile phone in use, the content of the SMS, the contact name and caller number. These types of information were not analysed in this thesis, but we will investigate the privacy vulnerabilities of these information types in order to raise user awareness as future work, as it is very important that the users become aware of what information is being collected from their fitness trackers, considering that a big amount of personal information is at risk.

Smart home devices and fitness trackers data sharing practices are also an aspect that users should be vigilant for. Reviewing the text of the privacy policies, an example that should raise the alertness of users is that in some cases when users grant access to a third party app to their smart home device or fitness tracker account, then the use of the account information will be governed by the third party's privacy policy, and not their device's policy. It is crucial for the users to be aware about such terms that are stated in the privacy policies, increasing the importance for application providers to disclose their privacy policies in a clear and easy to read manner, enabling the users to protect their privacy [145].

As privacy policies are used for the portrayal of the rights and responsibilities of both the user and the service provider in terms of data collection, processing and sharing, the GDPR forced the service providers to adapt their privacy policies content to the regulation's requirements, in order to provide the information required to the users. Even though the GDPR is aimed towards data controllers, the users are what the content is really about. The problem of the lack of proper communication from the side of service providers to the users in relation to data collection, processing and sharing practices in the privacy policies text is examined in the next chapter, where we aim to enhance the user awareness of smart home devices and fitness trackers regarding the protection of their privacy, by analysing the privacy policies of such devices in order for the users to be informed about the data collection, processing and sharing practices of the service provider that collects their data, as well as the potential risks that are present from the possible inferences.

Alexia Dini

Chapter 8

Extracting GDPR user rights and inference risks from privacy policy texts

Alexia Dini

Regulations and laws, such as the GDPR, require service providers to inform the users about their data collection and processing practices. The existing method used for the portrayal of the rights and responsibilities of both the user and the service provider in terms of data collection, processing and sharing, are the privacy policies, that depict the practices that an organisation or company follows when handling the personal data of its users. The GDPR explicitly defines eight distinct rights that all European citizens are entitled to and that service providers must respect through their data practices.

In this Chapter, we introduce *SpotAware*, an automated approach that: a) classifies the text of privacy policies from the domains of fitness trackers and smart homes, extracting information regarding the eight GDPR user rights addressed in the privacy policy, and b) classifies the text of privacy policies from the same domains extracting information about possible data inferences (e.g. location, health status) that can be drawn about the user based on the collected data as described in the text. The Chapter also presents the implementation of the presented approach in the PrivacyEnhAction web application, through which the users can better understand how their personal data are being collected and used by various smart home devices and fitness trackers and be informed abut the potential inferences that can be made about them based on the policy text.

8.1 Contributions

The aim of GDPR is to protect the users and their rights, which are recorded as the *Rights of the Data Subject* in Chapter 3 of the GDPR. Furthermore, GDPR Articles 12-14 designate that *data controllers must communicate any mandatory information or information relating to data processing to the user in a concise, transparent, intelligible and easily accessible form, using clear and plain language, as well as information necessary to ensure a fair and transparent processing* [93].

The work in this Chapter is driven by the problem of the lack of proper communication from the side of service providers to the users regarding their data practices and also the fact that users tend not to read the privacy policies of their devices. In this Chapter, we examine if the privacy policies of fitness trackers and smart home devices communicate the necessary information to the users with regards to data collection, processing and sharing, aiming to support the users of such devices to know which of the eight GDPR user rights are being addressed by the privacy policy of a device using supervised machine learning techniques. Furthermore, through the use of the same techniques, we intend to obtain the possible data inferences that can be drawn about the user from the privacy policy text and subsequently increase the user awareness about them. 133 privacy policies, which were available online at the time, were manually labelled with the above information.

In this chapter, we make the following contributions: (i) We provide a systematisation of inference groups that include possible inferences or conclusions that could be drawn about the users from privacy policy texts; (ii) We provide two annotated datasets of 133 privacy policies of smart home devices and fitness trackers for the two cases we study: a) extracting information regarding the eight GDPR user rights present in a privacy policy, and b) extracting information about possible data inferences that can be drawn about the user based on the collected data as described in the text of the privacy policy; (iii) We introduce a classification approach to see which GDPR rights are present in privacy policies and to indicate any data inference risks entailed in the policy text; (iv) We discuss our findings about which GDPR rights and inferences are more frequently found in privacy policies.

8.2 Approach overview

The two main goals of this part of this thesis are: (i) to analyse a given privacy policy (of a fitness tracker or a smart home device) in order to determine if and which of the eight GDPR user rights are being addressed by the particular policy, and (ii) to analyse a given privacy policy (of a fitness tracker or a smart home device) to detect what possible inferences could be drawn about the users according to the policy text under investigation. The implementation of this work was carried out using the Python programming language.

Alexia Dini

In this section, we describe the proposed approach, *SpotAware*, to analyse the privacy policies of fitness trackers and smart home devices. The approach can be replicated in other domains as well, using the available privacy policies of the preferred domain, or using the datasets created in the current work. SpotAware receives the URL of a privacy policy and then: (i) examines if and which of the eight GDPR user rights are addressed in the policy text, and (ii) scans the policy text to evaluate if any data inferences or conclusions can be extracted about the user from the policy text and the data collected from the device, according to the policy text.

The overview of our approach is shown in Figure 8.1, and it consists of three main steps. For the first part of this study, i.e. identifying the GDPR user rights addressed in privacy policies, which we call the *Rights Classification* component, the first step involves the creation of a list of GDPR-related terms that should be included in the privacy policy of a fitness tracker or a smart home device. Using as a basis the terms list from the work of Vanezi et al. [300], we have expanded and enriched that list with more terms, after manually inspecting 50 privacy policies of fitness trackers and smart home devices, as preliminary analysis (Table 8.1). We have created eight terms files, one for each GDPR user right, to use as input in the subsequent steps of the approach, which can be found at Appendix A.

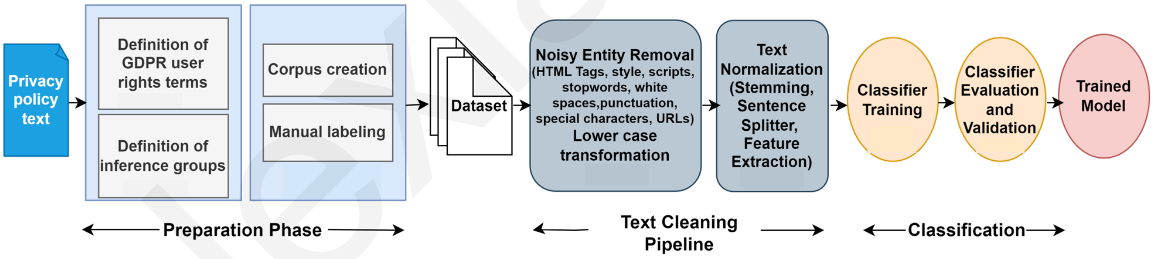

Privacy policy text
Definition of GDPR user rights terms
Definition of inference groups
Corpus creation
Manual labeling
Dataset
Noisy Entity Removal (HTML Tags, style, scripts, stopwords, white spaces,punctuation, special characters, URLs) Lower case transformation
Text Normalization (Stemming, Sentence Splitter, Feature Extraction)
Classifier Training
Classifier Evaluation and Validation
Trained Model
Preparation Phase
Text Cleaning Pipeline
Classification

Figure 8.1: Overview of our approach

For the second part of our work, i.e. the identification of any inferences emerging from privacy policies text, which we call the *Inference Detection* component, the first step was the thorough study and analysis of the text of 50 privacy policies of fitness trackers and smart home devices, in combination with the examination of the list of inferences we devised in Chapter 5 and can be seen in Table 5.1 and from the related literature. Through this analysis, we have compiled a list of seven categories of possible inferences.

The second step of our approach is concerned with the collection of privacy policies from the fitness trackers and smart home domains, the manual annotation of the collected privacy policies for each component and the preparation of a privacy policy corpus for each

Alexia Dini

Table 8.1: Number of terms in each GDPR user right list created for SpotAware

GDPR User Right	No of terms
Right to Be Informed	102 (all new)
Right of Access	30 (11 new)
Right to Rectification	33 (16 new)
Right to Erasure	29 (18 new)
Right to Restriction of Processing	24 (15 new)
Right to Data Portability	24 (5 new)
Right to Object	15 (9 new)
Right to Avoid Automated Decision Making	10 (all new)
Total number of terms	**267 (186 new)**

corresponding domain.

The third step includes the classification and prediction step. We have identified the privacy policy analysis task as a multi-label classification problem, as each sentence of text in a privacy policy may contain information relevant to any of the eight GDPR user rights or to any of the identified inference groups. Multi-label classification is a classification task that allows each data point to be assigned to more than one class at the same time [215] and is concerned with learning from a set of examples that are associated with a set of labels [290, 291]. Consequently, we create a corpus of 21,481 sentences from 133 privacy policies, using the labelling practices described in Section 8.3.4. We continue with the training of our classification models and the prediction step for new privacy policies, more details of which can be found in Section 8.3.7.

8.3 Methodology Steps

8.3.1 Privacy policies text collection

In order to create a list of privacy policies texts to be used in SpotAware, we searched online for the most popular commercial fitness trackers and smart home devices that we are interested in in this work, using the names of well-known vendors that provide such services (e.g. Garmin, Samsung, Xiaomi, Huawei for fitness trackers, and Vivint, Philips Hue, Arlo, Hive for smart home devices). We augmented our list by searching online to find other vendors available on the market that provide a privacy policy for the users, as through our research

Alexia Dini

we noticed that not all vendors offer a privacy policy. For the collection of privacy policies, the only rule followed was that the policy should be in English, and our quest led to a list of 75 fitness tracker and 135 smart home device privacy policies.

8.3.2 Definition of GDPR user rights terms list

For the Rights Classification component of SpotAware, in the first step of the methodology we have used as a basis the available terms list from the previous work of Vanezi et al. [300], where the authors provide a defined set of 89 GDPR terms for privacy policies in seven groups by examining a number of web platforms privacy policies. Out of these seven groups, six of them map to the corresponding user rights provisioned by the GDPR that we are interested in this work. The eight GDPR user rights are the following (the previous work does not include terms for the first and the last rights of the list):

1. The Right to Be Informed (Art. 13, 14)
2. The Right of Access (Art. 15)
3. The Right to Rectification (Art. 16)
4. The Right to Erasure (Art. 17)
5. The Right to Restriction of Processing (Art. 18)
6. The Right to Data Portability (Art. 20)
7. The Right to Object (Art. 21)
8. The Right to Avoid Automated Decision-Making (Art. 22)

Then we studied the relevant GDPR articles along with the relevant literature [15,44,113, 114,250,266] in order to understand how each right should be disclosed in a privacy policy and what kind of information should be provided in order to address a user right. Following this practice, we examined the text of 50 of the available privacy policies and extracted the terms that should be included for addressing each right separately, paying attention to the variations in how each of these terms were expressed in different policies, eventually reaching the final set of terms for each user right, expanding the list provided in [300]. This process was performed by one person and another person verified the findings, i.e. that the chosen terms were relevant to each of the eight rights considering the nature of each right. The number of terms that were collected can be found in Table 8.1. A subset of the final list of the GDPR user rights privacy policy terms referring to the Right to Avoid Automated Decision Making can be found in Table 8.2. The lists of terms for all the GDPR user rights can be found in Appendix A.

Alexia Dini

Table 8.2: Example of GDPR User Rights Terms List

Right to Avoid Automated Decision Making
• Automated Decision-Making, Including Profiling • Object to a decision based solely on automated processing, including profiling • Object to automated decision-making • Objecting to automated decision making and profiling • Right not to be subject to a decision based solely on automated processing • Right not to be subject to a decision which is based solely on automated processing • Right to object to automated decision making • Right to refuse to be subjected to automated decision making, including profiling • Rights related to automated decision making including profiling • Automated Decision-Making

8.3.3 Definition of inference groups

The first step of the methodology for the Inference Detection component of SpotAware involves determining the possible data inferences that can be extracted about the user through the privacy policies text. Since this work concerns fitness trackers and smart home devices, as a first step we have relied on the inferences we have already defined for the identification of a group of inferences emerging from the privacy policies under study. At a second level, we inspected carefully the content of 60 of the available privacy policies for any text indicating or hinting the prospect of deducing any assumptions or conclusions about the user. Using the above process, we devised a final list of seven inferences groups, which are summarised in Table 8.3. In the table, we have included some examples of text sentences from policies that indicate how such an inference could be possible.

Table 8.3: List of inference groups deducted from privacy policies

No	Inference group	Examples	Brand	Type
1	**Profiling/Identifying User**	*"This information is the data you give during registration: email, name, age, height, sex, training background and location"*	Polar	FT

Continued on next page

Alexia Dini

Table 8.3: List of inference groups deducted from privacy policies (Continued)

No	Inference group	Examples	Brand	Type
		“Personal Information is information that allows someone to identify or contact you, as well as any other non-public information about you that is associated with or linked to any of the foregoing, and can include, for example, an individual’s first and last name, address, phone number, e-mail address, IP address, location or other personally identifiable details."	Lifx	SHD
2	**Location/Occupancy**	*“When you access certain location-based services (such as perform searches, use navigation software, or view the weather for a specific location), we will collect, use, and process the approximate or precise location of your device."*	Huawei	FT
		“For example, this includes information that smart devices you connect to your Hive Hub collect about rooms’ temperatures, temperature settings, heating schedules, lighting use and schedules, when contact sensors show doors or windows are open or closed, when motion sensors detect movement, water flow events and water temperature, when smartplugs are on or off and video and audio recordings from your monitoring devices."	Hive	SHD
3	**Inferences from combining data from various sources**	*“If you choose to sign on using these services, Motorola may collect certain information from your social media account, including your public profile, email address, age/date of birth, contact lists, interests, likes, and current city."*	Motorola	FT
		“We may also receive information about you from social media platforms, for instance, when you interact with us on social media."	Sonos	SHD
4	**Inferences from disclosing data to third parties**	*“You hereby agree that the company deals with and discloses personal data and SPDI to affiliates (communications, social media, technology and cloud business), third party service providers (hereinafter defined) for the purposes set out in this privacy policy*."	Boat	FT
		“Professional advisors and others: we may share your data with other parties including professional advisors, such as banks, insurance companies, auditors, lawyers, accountants, other professional advisors."	Philips Hue	SHD
5	**Inferences leading to targeted advertising**	*“We use this information to analyse your preference, habit and location, etc. so as to provide more tailored services to you (e.g. accurately record your movement trajectory and provide advertising and promotion information that better meets your needs)"*	Zeblaze	FT

Continued on next page

Alexa Dini

Table 8.3: List of inference groups deducted from privacy policies (Continued)

No	Inference group	Examples	Brand	Type
		“We also collect information about you from other third parties, for example marketing companies and data brokers, in order to better understand your interests and deliver you with more tailored Services and advertising."	LG	SHD
6	**Inferences about health status**	*“Inferences drawn from any of the above, including the number of calories you burned, distance you travelled, sleep insights, and personalised exercise and activity goals."*	Fitbit	FT
		“Genetic, physiological, behavioural, and biological characteristics, or activity patterns used to extract a template or other identifier or identifying information, such as, fingerprints, face-prints, and voice-prints, iris or retina scans, keystroke, gait, or other physical patterns, and sleep, health, or exercise data."	GE Appliances	SHD
7	**Inferences leading to sleep patterns extraction**	*“With the aim of enabling you to understand and to improve your sleeping habits, some Devices collect sleep start time, sleep end time, the time you go to bed, and the time you wake up."*	Misfit	FT
		“Genetic, physiological, behavioural, and biological characteristics, or activity patterns used to extract a template or other identifier or identifying information, such as, fingerprints, face-prints, and voice-prints, iris or retina scans, keystroke, gait, or other physical patterns, and sleep, health, or exercise data."	GE Appliances	SHD

Alexia Dini

In the following paragraphs, we analyse how each inference group was conceived and what kind of information is systematised under each group. Table 8.4 summarises various inferences as found in recent literature that assisted in the definition of the seven inference groups.

Group 1: Profiling/Identifying User. Under this category goes any information that can be used to identify the user. Data, such as name, date of birth, gender, weight, or height, are personal information that relate to an individual and could be used to identify the specific individual alone or in combination with other identifiers, such as an IP address. Profiling is the use of personal data to evaluate specific attitudes related to the individual, as for example personal preferences, interests, reliability, behaviour, sexual orientation, health status,

Table 8.4: Inferences found in literature

Group	Inference	DM	Cit.	Year
1	**Profiling/Identifying User**			
	Identity, personal traits, activities, habits, preferences, sexual orientation, health status, financial situation	SH	[178]	2022
	Identity theft	FT	[34]	2019
	Religion	FT	[302]	2021
	Activities	FT	[283]	2015
	Smoking	FT	[275]	2014
2	**Location/Occupancy**			
	Home address, points of interest, preferences, habits, religion, health	FT	[162]	2018
	User typical routes	FT	[323]	2015
	Occupancy	SHD	[96]	2017
3	**Inferences from combining data from various sources**			
	Unfair treatment, discrimination	FT	[160]	2020
	Location	SHD	[161]	2021
4	**Inferences from disclosing data to third parties**			
	Discrimination	FT	[170]	2014
5	**Inferences leading to targeted advertising**			
	User interests	SHD	[140]	2022
	User interests	FT	[87]	2021
6	**Inferences about health status**			
	User health status	FT	[83]	2023
	Drug or alcohol abuse	FT	[302]	2021
	Physical or mental health, level of intoxication	SHD	[165]	2020
7	**Inferences leading to sleep patterns extraction**			
	Lack of sleep, work performance	SHD	[224]	2014
	Health issues	FT	[83]	2023

Abbrev. used: DM = Domain; SHD=Smart Home Devices; FT=Fitness Trackers

Alexia Dini

financial situation, etc. even when data is collected anonymously. The huge amounts of data collected by devices increase the risk for profiling, upsurging simultaneously the danger that information about other parts of people's lives are being disclosed as well. A pertinent example is a smart thermostat with a temperature zone control that knows exactly which person is where in the house and when, aggregating sensitive user information that was previously inaccessible [178]. Asus privacy policy [17], regarding the collection of personal data, states that the following information is collected: "*Your age, gender, height, weight, body temperature, heart rate, blood pressure, movement of belly as well as certain data about your daily activities, for example, your step taken, calories burned, sleep patterns and diary records when you use our healthcare products and services*".

Group 2: Location/Occupancy. In this group, we have included the location and occupancy categories. The location history that can be tracked from fitness tracker data can reveal the home address and points of interest of the users [162] and as such any information that can reveal the user's geographical location goes under this category. In most cases users worry about location data [302] and fitness trackers request the collection of location data as mandatory. Location data can reveal a lot about the users, including where they sleep, work, socialise or seek medical treatment, as well as sensitive personal aspects, like habits, preferences or religion. The Fitbit privacy policy indicates that: "*Your device collects data to estimate a variety of metrics like the number of steps you take, your distance travelled, calories burned, weight, heart rate, sleep stages, active minutes, and location*". We have also categorised occupancy under this group, as smart home devices collect data that when analysed can reveal the occupancy status of the users' premises [161]. An example taken from Ecobee: "*Some device models may include additional types of data such motion sensing (i.e., occupancy sensing)*" [89].

Group 3: Inferences from combining data from various sources. The advancements and popularity of wearable and smart home devices that provide the means to users to record every aspect of their daily life, personality, behaviour, habits and location, bring along many privacy risks associated with the processing of their personal information. Such data can be combined with other datasets to make inferences about the users, creating further privacy risks including the risk of unfair treatment based on data about a person's assumed or actual health status [160]. For example, when smart home data are combined with data from social media like Facebook, it is possible to predict the users location at a specific time using

Alexia Dini

knowledge released by friends [161]. The Cubot privacy policy states the following: "*We may legally obtain information about you from other sources, and combine this data with information we already have about you*" [72].

Group 4: Inferences from disclosing data to third parties. The processing of data, and in particular of health-related data is illicit under the GDPR. The use of inferences obtained from fitness trackers and smart home devices data can assist in the creation of user profiles when such data are disclosed to third parties and inevitably could threaten the user privacy. Attentive third parties, like insurance companies, banks or employers, could use this information for decision-making processes related to insurance rates, loans, promotions, etc. [83]. The Hive privacy policy states: "*Companies involved in providing you with insurance: if you have bought or been given Hive devices alongside a home insurance policy, whether from British Gas or another third party insurance provider, companies involved in providing you with insurance, or giving you quotes for insurance, may use the fact that you have these Hive devices, along with information about your usage of those devices, in assessing their pricing and policy terms, and to assess your eligibility to claim under your insurance policy*" [132].

Group 5: Inferences leading to targeted advertising. Iqbal et al. in their work [140] have reported that data and user interactions with smart speakers are collected by Amazon and third parties and shared with advertising partners to infer the users' interests and provide them targeted advertising both on the platform they are using as well as on the web. As voice input data is typically stored on cloud servers for processing, sharing with third parties is imperceptible to the users. According to Aksu et al. [4], the IoT gives the opportunity to marketing companies to grow their targeted groups in order to deliver their ads adapted to the needs of the users' profiles, created through their collected data. Marketing companies exploit user data from smart home devices for the construction of user profiles, to predict user behaviour, aiming to boost the success of their advertising programs [227]. The Furbo privacy policy states: "*As permitted by applicable law, we process such information to better understand you, to maintain and improve the accuracy of the information we store about you, to deliver targeted marketing based on your interests and preferences, and to better promote or optimise our Services*" [108].

Group 6: Inferences about health status. As fitness trackers log the activities performed by the users on a daily basis, insights about these types of information could provide the abil-

Alexia Dini

ity to make inferences about the health status of the users, based on the amount of activity they perform [83]. Furthermore, it has been shown in the literature that through the use of machine learning techniques it is possible to infer sensitive information about the users from their fitness tracker data, such as drug or alcohol consumption [302]. The MyKronoz privacy policy states: “*When you synchronise your MyKronoz device by means of a MyKronoz application, you transmit automatically to our servers activities data regarding your physical shape and your health, namely, the number of steps made, the number of calories burnt, the number of hours slept or your heart rate*" [203]. In the case of smart home devices that collect voice input data, such as smart speakers, the voice and tone of a user can disclose information related to the physical or mental health of the user and the level of intoxication, through advanced data analytics techniques [165].

Group 7: Inferences leading to sleep patterns extraction. Data from smart home devices can reveal a lot of information about the home owners, including insights about how sleep-deprived a person may be [224]. Furthermore, research has shown that data from wearables can be used to evaluate the sleep patterns of the users, sleep fragmentation and sleep efficiency [164]. As sleep is a vital element in people’s well-being, lack of or bad quality of sleep has been connected with health issues [83]. The Alcatel privacy policy states: “*Information relating to your health status and details relating to your fitness and exercise information (such as your height and weight, body temperature, heart rate, sleep quality, walking, running and biking tracker information) to the extent that you provide us with such information or otherwise consent to our collecting*" [9].

Alexia Dini

8.3.4 Privacy policy corpora preparation

For the second step of the methodology followed in SpotAware we use the list of 75 fitness trackers and 135 smart home device privacy policies collected in the fist step.

Extraction and pre-processing:. In order to be able to annotate the content of the privacy policies collected, first we extracted the text of each policy. We have used the *BeautifulSoup* Python library [242], which transforms a complex HTML document into a parsed tree of Python objects, that can be used to extract data from HTML. Given the URL of a specific privacy policy, we parsed the content into a BeautifulSoup object, iterated over the data to remove the tags from the document using the *decompose()* method and by using the *stripped_strings()* method we retrieved the tag content. The *decompose()* method removes a tag from the tree of a given HTML document, and then completely destroys it and its contents, while the *stripped_strings()* method is used to remove any whitespace at the beginning and end of strings. Following this process, we parsed the content from the given privacy policy URL, removing all style, scripts, and HTML tags, retrieving only the actual privacy policy main content of the page.

Then, the policy text was processed using the Python *sentence-splitter* module we have adopted from [155], which enables the splitting of text paragraphs into sentences. The module uses punctuation and capitalisation clues to split paragraphs into a newline-separated string with one sentence per line. The result of this process (ran for each privacy policy) was saved in a separate csv file, and then all the files were merged to be used in the labelling process. It was observed that a number of privacy policies were not available online at that specific time as the URL would not open, reducing the number of available policies to 66 fitness tracker and 132 smart home devices privacy policies. As the manual annotation of the policies was performed by the authors and due to resource limitations, it was decided to label and use the similar number of policies from both sectors, and as such, 66 fitness tracker and 67 smart home devices privacy policies were labelled. In total, our fitness tracker corpus contains 10,622 sentences and the smart home devices corpus 10,859 sentences. More details about the annotation process are provided in Section 8.3.2, while the statistics of our corpora can be seen in Table 8.5. The annotated datasets of the combined privacy policies from both fitness trackers and smart home devices are available at the Zenodo repository[1].

Semi-Automated Annotation:. For the Rights Classification component, after defining the lists of GDPR terms (see Section 8.3.2), we proceeded with preparing two sets of privacy policies: one for fitness trackers and one for smart home devices. As a first step towards the labelling of the 133 files and in order to assist the labelling process, we ran a script for the automatic labelling of the files where an exact match of a term was searched for. After this initial preliminary automated labelling process, we proceeded with the manual labelling. Here, three experts had to read the content of every policy separately and perform manual annotation when required. Each privacy policy was read and annotated by one of the experts involved in the process (the whole set of policies was divided among the experts involved), but a subset of the texts was examined by a second expert (30 texts in total were examined by two experts in total). This was performed in order to ensure that all experts

[1] https://doi.org/10.5281/zenodo.7934945

Alexia Dini

Table 8.5: Statistics of privacy policy corpora

Context	No of Pr. Policies	No of Sentences	RCC-An.Sentences No	IDC-An.Sentences No
SHD	67	10,859	4,765	3,103
FT	66	10,622	4,716	1,035
Total	**133**	**21,481**	**9,481**	**4,138**

Abbrev. for Context: SHD = Smart Home Devices; FT = Fitness Trackers

Other Abbrev. Used: RCC = Rights Classification Component; IDC = Inference Detection Component

agree on the context the policy sentences address regarding the particular user right(s). Cases of disagreement were few and consensus was reached among all experts that continued with annotating the rest of their dataset part after reaching consensus on the 30 texts examined by two experts to ensure that the same rationale was used in the labelling process. On average, 35 minutes were required by each expert for the annotation of each privacy policy.

For the Inference Detection component, and in order to prepare the privacy policy corpora for this part of work, we used the listing of 66 fitness trackers and 67 smart home devices privacy policies created earlier. Similar to before, the task was to manually annotate the available policies using the 7 inference groups described in Section 8.3.3. Using the inference groups along with the identified inferences from the literature, we proceeded with studying the content of every privacy policy separately and manually annotating it when it was observed that an inference was apparent. In a similar manner, each privacy policy had to be read by one of the experts but initially, a subset of 30 texts in total was examined by two of the experts, following the exact same process as for the case of the GDPR user rights. Only a few cases of disagreement were encountered and experts agreed on them, before proceeding with the labelling of the rest of the texts.

Alexia Dini

8.3.5 Data pre-processing

In SpotAware, in order to be able to check the URL of a given online privacy policy and also for the preparation of the privacy policy corpora, we needed to extract the plain text of the policy. In order to accomplish this, we have used the *BeautifulSoup* Python library [242], parsing the content from the given URL, removing all style, scripts, and HTML tags, retrieving only the actual privacy policy main content of the page. Then, the policy text was processed by the Python *sentence-splitter* module [155], which enables the splitting of text paragraphs into sentences.

For the pre-processing of the text sentences, we have used the Natural Language Toolkit

(NLTK) Python package for natural language processing [278]. Punctuation and special characters were removed, the text was transformed to lower-case and stop-words were removed (using the default set of stop-words from the *nltk.corpus* Python library). Stemming was applied afterwards. Next, for the feature extraction from our policy text, we used *term-frequency inverse document frequency* (TF-IDF), an algorithm for numerical statistics that shows the importance of a word in a document in a corpus, creating a vocabulary of all the words in the corpora. *TF* counts the occurrences of each term in each document, where *tf(t,d)* is the number of times the term *t* occurs in document *d* (Equation 8.1 [169]). *IDF* computes how common a word is in the corpus (Equation 8.2 [169]). Based on the that, TF-IDF is calculated as shown in Equation 8.3 [169]. The importance of a term, i.e. its TF-IDF score, increases proportionately with the number of times it appears in the corpora, while it declines in inverse proportion to the frequency of its occurrence, disqualifying ordinary terms to be selected as important features [97].

$$tf(t,d) = \frac{f_d(t)}{\max f_d(w) : w \in d} \tag{8.1}$$

$$idf(t,D) = \ln\left(\frac{|D|}{|d\epsilon D : t\epsilon d|}\right) \tag{8.2}$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{8.3}$$

8.3.6 SpotAware multi-label classification

For the purposes of this work, the privacy policy analysis task is modeled as a multi-label classification problem. In multi-label learning, the input to the learning algorithm consists of a set E of N classified examples { (x_1, Y_1),. . . , (x_N,Y_N) }, where each example E_i =(x_i, Y_i) is associated with a set of labels Y_i, where $Y_i \subseteq$ L, $Y_i \neq \emptyset$ and L = {y_1, y_2,. . . , y_q}. Here, the multi-label algorithm generates a classifier H which, given a new example x, H(x) predicts the corresponding multi-label Y [59].

For the Rights Classification component, this is because each sentence of text in a privacy policy may contain information relevant to one or more of the eight GDPR user rights, while for the Inference Detection component, a sentence of text in a privacy policy may enclose information that could lead to one or more of the seven inference groups. In order to identify which is most appropriate classifier for the particular dataset attributes of this study we

Alexia Dini

evaluated the following multi-label classifiers using the *scikit-multilearn* library provided by the *scikit-learn* high level framework:

- MLkNN: This algorithm is derived from the kNN algorithm. Initially, the k-Nearest Neighbours of the test instance is established, and then by using the MAP rule (Maximum A Posteriori) and the k value, the set of labels for the test instance are determined. This classifier was selected as it was the first method that was specifically created for multi-label classification and also because it is one of the most broadly cited algorithm adaptation methods in the literature [57].
- BinaryRelevance: This classifier works by transforming a multi-label classification task with L labels into L binary classification tasks [337]. In BinaryRelevance classifier each target variable $(y_1, y_2, \ldots, y_n)$ is treated independently. It was chosen as it is the baseline algorithm for multi-label classification [184].
- BRkNN-a: The BRkNN-a classifier, an extension to the BRkNN classifier, returns the labels that give the highest score even if these labels are lower than the threshold, as in most multi-label datasets it is not common to have an empty set of labels. Along with BRkNN-a, the BRkNN-b classifier is another extension to BRkNN that reduces the cardinality of the labels between the predicted and the actual label sets. We have rejected this classifier as it is more suitable for datasets with high cardinality [53], considering the nature of our datasets that have low cardinality. The BRkNN-a classifier was selected as it is a popular lazy learning method that is simple and efficient at the same time [59].
- ClassifierChain (CC): The ClassifierChain classifier is similar to the BinaryRelevance, but the target variables are not fully independent. A Bayesian conditioned chain of per label classifiers is constructed, exploiting any possible correlations between labels, and thus assisting in the improvement of performance. This classifier was chosen as it is one of the state-of-the art currently used learning algorithms for multi-label classification [57].

8.3.7 Multi-label classification evaluation

In multi-label classification, partially correct predictions should also be acknowledged in the evaluation of a model. Different performance measures can be used for the assessment of multi-label classification problems. In this work, we compare the results of the algorithms used based on a range of evaluation metrics for multi-label classification, such as the Exact

Alexia Dini

Match Ratio, the Hamming Loss, the Micro F-measure and the Macro F-measure. The Exact Match Ratio indicates the percentage of samples that have all their labels classified correctly (8.4). Hamming Loss reports how many times on average, the relevance of an example to a class label is incorrectly predicted, as defined in Equation 8.5, where I is the indicator function. Ideally, the hamming loss should be 0, implying no error, but the smaller the value of hamming loss, the better the performance of the learning algorithm. In order to evaluate the performance of a model thoroughly, the precision and recall metrics should be examined. Precision is the proportion of predicted correct labels to the total number of actual labels, averaged over all instances (Equation 8.6). Recall is the proportion of predicted correct labels to the total number of actual labels, averaged over all instances(Equation 8.7). The F1 Score considers both of them as can be seen in Equation 8.8. The Macro F-measure is the macro-averaged F1 score and is calculated using the arithmetic mean of all the per-class F1 scores, while the Micro F-measure calculates a global average F1 score by counting the sums of the True Positives (TP), False Negatives (FN), and False Positives (FP).

$$\text{Exact Match Ratio, MR} = \frac{1}{n}\sum_{i=1}^{n} I\left(y_i = \hat{y}_i\right) \tag{8.4}$$

$$\text{Hamming Loss, } HL = \frac{1}{kn}\sum_{i=1}^{n}\sum_{l=1}^{k}\left[I\left(l \in Z_i \wedge l \notin Y_i\right) + I\left(l \notin Z_i \wedge l \in Y_i\right)\right] \tag{8.5}$$

$$\text{Precision, } P = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{8.6}$$

$$\text{Recall, } R = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{8.7}$$

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2\,|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \tag{8.8}$$

8.4 Experimental Setup

All the algorithms were implemented in Python 3.9.0, while the experiments were carried out on an i5 2.5 GHz machine with 4GB RAM. The experiments were run on the five algorithms as discussed in Section 8.3.6.

We conducted the following experiments on the classification algorithms selected using the available datasets as follows: (i) Experiment 1: Fitness Tracker dataset, (ii) Experiment 2:

Alexia Dini

Table 8.6: SpotAware Rights Classification component - Classification results

Data	Methods	EMR	HL	MiF1	MaF1	MiR	MaR	MiP	MaP
FT	**MLkNN**	0.7085	0.0389	0.6386	0.5952	0.5787	0.5020	0.7122	0.7851
	BRkNN-a	**0.9446**	**0.0115**	**0.9548**	0.6778	**0.9284**	*0.5547*	**0.9826**	**0.9410**
	BR	*0.8107*	*0.0253*	*0.7765*	*0.6853*	*0.7407*	0.5532	*0.8160*	*0.9364*
	CC-SVC	0.8102	0.0254	0.7757	**0.6999**	*0.7407*	**0.5792**	0.8143	0.9172
	CC-MNB	0.7538	0.0337	0.6709	0.0916	0.5782	0.0846	0.7988	0.0998
SHD	**MLkNN**	0.7354	0.0355	0.7269	0.5907	0.7426	0.5162	0.7118	0.7329
	BRkNN-a	0.7336	0.0362	0.7142	0.3906	0.7097	0.2882	0.7187	0.8476
	BR	*0.9564*	**0.0087**	**0.9659**	**0.7776**	**0.9513**	**0.6605**	**0.9808**	**0.9786**
	CC-SVC	**0.9585**	*0.0097*	*0.9622*	*0.7483*	*0.9453*	*0.6437*	*0.9797*	*0.9721*
	CC-MNB	0.9202	0.0219	0.9139	0.1632	0.887	0.1488	0.9405	0.3016
Comb.	**MLkNN**	0.6972	0.0400	0.6367	0.6072	0.5753	0.5091	0.7127	0.7627
	BRkNN-a	0.6452	0.0466	0.4216	0.4709	0.2783	0.3317	**0.8685**	**0.9159**
	BR	**0.8027**	**0.0261**	**0.7766**	**0.7246**	**0.7450**	**0.6109**	*0.8110*	*0.9069*
	CC-SVC	**0.8027**	*0.0264*	*0.7733*	*0.6983*	*0.7398*	*0.5787*	0.8100	0.9042
	CC-MNB	*0.7519*	0.0338	0.6911	0.1087	0.6203	0.0968	0.7800	0.3476

Abbr. for Methods: BR=BinaryRelevance; CC-SVC=ClassifierChain with SVC as base-learner; CC-MNB=ClassifierChain with MultinomialNB as base-learner

Abbrev. for Evaluation metrics: EMR=Exact Match Ratio; HL=Hamming Loss; MiF1=Micro F1; MaF1=Macro F1; MiR=Micro Recall; MaR=Macro Recall; MiP=Micro Precision; MaP=Macro Precision

Smart Home Devices dataset, and (iii) Experiment 3: Combined Fitness Tracker and Smart Home Devices dataset, for the Rights Classification component and the Inference Detection component. As the datasets were already split into training and test sets with a 60:40 split ratio, we ran each classifier on the training set and report its performance on the test set. For all the experiments, we used ten-fold cross validation. For the lazy classification classifiers MLkNN and BRkNN-a, we varied the value of Nearest-Neighbors K in order to find the optimal value for each algorithm, which was found to be $K = 10$ for both cases, while the smoothing parameter was kept to the default value of 1. The BinaryRelevance classifier was used with Support Vector Machines (SVM) as base learner, as it has been shown to be the best classifier for this method [331].

8.4.1 SpotAware overall results

The experimental classification results of the algorithms for the Rights Classification component can be seen in Table 8.6, while the classification results for the Inference Detection component can be found in Table 8.7. These tables present the average values of the evaluation metrics, along with the micro-macro precision and recall values, while the best values are highlighted in bold and the second best in italics.

Alexia Dini

Table 8.7: SpotAware Inference Detection component - Classification results

Data	Methods	EMR	HL	MiF1	MaF1	MiR	MaR	MiP	MaP
FT	**MLkNN**	0.8950	0.0169	0.5817	**0.4813**	**0.4971**	**0.4295**	0.7010	0.5702
	BRkNN-a	0.8721	0.0220	0.1542	0.0730	0.0845	0.0407	0.8805	0.5719
	BR	**0.9150**	**0.0138**	**0.6148**	*0.4323*	*0.4641*	*0.3317*	*0.9101*	**0.7715**
	CC-SVC	*0.9140*	*0.0142*	*0.5988*	0.4071	0.4469	0.3061	0.9069	*0.7630*
	CC-MNB	0.8750	0.0217	0.1598	0.0701	0.0873	0.0393	**0.9384**	0.3904
SHD	**MLkNN**	0.8655	*0.0221*	0.5280	0.4228	0.4304	**0.3342**	0.6830	0.6443
	BRkNN-a	0.8573	0.0237	0.3412	0.1431	0.2128	0.0919	0.8605	0.5379
	BR	*0.8928*	**0.0173**	**0.6094**	*0.4258*	**0.4684**	0.3164	*0.8716*	**0.8022**
	CC-SVC	**0.8940**	**0.0173**	*0.6088*	**0.4285**	*0.4673*	*0.3172*	**0.8733**	*0.8005*
	CC-MNB	0.8439	0.0262	0.2058	0.0740	0.1177	0.0443	0.8181	0.2932
Comb.	**MLkNN**	0.8589	0.0240	0.2001	0.0807	0.1128	0.0466	**0.8800**	0.3692
	BRkNN-a	0.8570	0.0249	0.1579	0.1298	0.0878	0.0752	0.7784	0.5547
	BR	**0.9059**	**0.0152**	**0.6442**	**0.5237**	**0.5150**	**0.4282**	0.8597	*0.8603*
	CC-SVC	*0.9057*	*0.0153*	*0.6387*	*0.5207*	*0.5080*	*0.4241*	*0.8599*	**0.8644**
	CC-MNB	0.8589	0.0240	0.2001	0.0807	0.0112	0.0466	**0.8800**	0.3692

Rights Classification:. For the Rights Classification component of the SpotAware approach, we observe the following. For the fitness trackers dataset, the BRkNN-a algorithm outperforms the other algorithms producing better scores in almost all evaluation metrics, with 94% exact match ratio, 0.0115 Hamming Loss and 95% Micro F1 score. The best Macro F1 score, which is a more suitable evaluation metric for the Rights Classification component as our datasets are imbalanced, is given by ClassifierChain with SVC as base learner with a score of 69,99%. We define our Rights Classification component datasets as imbalanced as in the first step of our methodology, described in Section 8.3.2, in the definition of the GDPR user rights term list, it was obvious from the privacy policies text that the first GDPR user right, i.e. *The Right to Be Informed*, is being addressed in a wider variety of terms in the text of privacy policies and as such, the manual labelling of the privacy policies for the Rights Classification component resulted in an uneven distribution of observations between the target classes, where the class label for the first GDPR user right is having a much higher number of observations compared to the rest. Furthermore, as our dataset has low cardinality, the use of the BRkNN-a algorithm has proved to solve this problem [53]. The BinaryRelevance and ClassifierChain with SVC as base-learner also performed very well, delivering similar evaluation metrics scores between them. This is quite expected as the ClassifierChain classifier is built on the strategy of the BinaryRelevance classifier, but compared to the BinaryRelevance, the ClassifierChain algorithm can also pick up the interconnections between each pair of labels [277]. We can see a slightly improved performance of the ClassifierChain

Alexia Dini

algorithm over the BinaryRelevance approach in the F1 score. For the smart home devices dataset and the combined approach experiment, the BinaryRelevance classifier has the best overall performance. When performance is measured by the Hamming Loss metric, we can see that BinaryRelevance achieved the best results in two out of three datasets. This outcome is expected since BinaryRelevance optimises the Hamming Loss measure when a proper base learner is being used [184]. We have chosen to use Hamming Loss as our main evaluation metric as it is probably the most widely used loss function in multi-label classification and also because it considers both prediction and missing errors [175].

Inference Detection:. For the Inference Detection component of the SpotAware approach, observing the results we can see that the BinaryRelevance classifier has the best overall performance in all the datasets, followed very closely by the ClassifierChain classifier with SVC as a base learner. From the results, it is obvious that BinaryRelevance shows the best Hamming Loss score for the fitness trackers dataset, followed by ClassifierChain-SVC. For the Smart Home Devices dataset and the combined dataset, both these classifiers produce the best Hamming Loss score, with a value of 0.0173 for both classifiers in the first experiment, and 0.0152 and 0.0153 in the second respectively.

Alexia Dini

8.4.2 SpotAware results per user right and inference group

Tables 8.8 and 8.9 show the performance results per class of the best performing classifiers BinaryRelevance and ClassifierChain with SVC as base learner for the combined dataset for each classification category in the Rights Classification and Inference Detection components.

Rights Classification:. In the Rights Classification component of SpotAware, with the BinaryRelevance classifier the recall rates range from 51% to 77%, the precision rates from 66% to 97%, while the overall F1 score is 69.77%. For the ClassifierChain the corresponding rates are 45% to 77% for recall, 66% to 95% for precision, and the overall F1 score is 65.58%. The ClassifierChain classifier works best on the first GDPR user right, *TThe Right to Be Informed*, with an F1 score of 79.67%, while the BinaryRelevance classifier also performs very well on that right giving an F1 score of 78.5%. Both classifiers had the worst F1 scores on the eighth GDPR user right, *Right to Avoid Automated Decision Making*, with Binary relevance giving an overall score of 59.63% and ClassifierChain a score of 56.68%. This demonstrates the case that recognition accuracy is proportional to the dataset size [45]. As the first GDPR user right is very broad, it refers to information that must be provided by

Table 8.8: RCC - BR and CC prediction performance for combined dataset

GDPR Right	**BinaryRelevance**			**CC-SVC**		
	Prec.%	Rec.%	F1%	Prec.%	Rec.%	F1%
Right to Be Informed	79.80	**77.75**	78.50	79.66	77.50	79.67
Right of Access	87.20	50.76	64.00	89.20	50.20	64.00
Right to Rectification	92.68	56.68	70.22	90.84	50.00	64.50
Right to Erasure	83.38	51.73	65.00	93.50	45.67	61.34
Right to Restriction of Processing	94.10	56.50	70.37	93.17	47.60	62.66
Right to Data Portability	**97.40**	51.97	67.33	95.40	47.20	60.50
Right to Object	96.36	73.30	**83.10**	92.80	63.84	75.33
Right to Avoid Aut. Decision-Making	65.78	54.52	59.63	65.82	49.76	56.68
Overall	87.09	59.15	69.77	87.55	53.97	65.58

the data controllers when they collect personal data directly from data subjects, including information about what kind of data they process, why the data controller needs that data, the legal basis for processing and purposes of processing, the legitimate interests of the processor and third parties, any recipients of personal data, or the explanation of the right to withdraw consent and to complain to the relevant supervisory authority, among others. To that end, there is a plethora of statements in privacy policies fit for the purposes of satisfying these requirements. On the other hand, it has been observed by the annotators that the *Right to Avoid Automated Decision-Making* has not been fully adapted by the providers of fitness trackers and smart home devices, as a relatively small number of occurrences of the ten relevant terms defined for this user right have been found in the 133 privacy policies under study.

Alexia Dini

Inference Detection:. For the Inference Detection component of the SpotAware approach, the BinaryRelevance classifier gives an overall F1 score of 51.85% and the ClassifierChain classifier an F1 score of 48.47%. Looking at the individual scores per inference groups we observe that there is one case where the F1 score is very low (*Inferences from disclosing data to third parties*) with 4% in BinaryRelevance and 4.84% in ClassifierChain. This is not an unexpected result, as the task to distinguish between sentences containing text regarding the rightful disclosure of data to third parties, and sentences containing text involving the unlawful disclosure of data to other parties, is not straightforward. This is a result of the GDPR requirement that data controllers must include clauses in their policies that reflect

Table 8.9: IDC - BR and CC prediction performance for combined dataset

Inference Group	**BinaryRelevance**			**CC-SVC**		
	Prec.%	Rec.%	F1%	Prec.%	Rec.%	F1%
Profiling/Identifying User	84.33	48.50	61.66	86.00	46.67	60.50
Location/Occupancy	87.66	66.33	75.66	**87.67**	59.50	70.84
Inf. from combining data from various sources	80.00	18.83	30.33	79.17	15.50	25.83
Inf. from disclosing data to 3rd parties	61.16	2.00	4.00	66.67	2.33	4.84
Inf. leading to targeted advertising	86.50	65.50	75.33	87.26	64.00	73.84
Inf. about health status	82.00	24.50	37.17	85.84	22.00	34.44
Inf. leading to sleep patterns extraction	97.33	**66.67**	**78.83**	77.20	55.83	69.00
Overall	82.71	41.76	51.85	81.40	37.98	48.47

third-party disclosures, therefore it is not easy to distinguish when data disclosure to third parties is performed for the legitimate interests of the data controller following the legal basis for processing, or when it involves an unlawful data disclosure. Furthermore, most privacy policies do not provide adequate information to users about their third party affiliates. Looking further at the results, we can see that for the *Inferences from combining data from various sources* inference group, the BinaryRelevance F1 score is 30.33% and the ClassifierChain classifier F1 score is 25.83%, both of them being below 50%. This is probably due to the fact that in most privacy policies information accumulated through the combination of information collected by various sources is termed as aggregated and non-personal by the providers, and to that end it is not easy to differentiate between text sentences referring to personal and non-personal combined data. For the *Inferences about health status* inference group both classifiers also scored low F1 values; in particular, the BinaryRelevance classifier returned 37.17%, while the ClassifierChain returned 34.44%. For the particular case, both models seem to miss a lot of sentences containing text relevant to this inference group we created (as described in Section 8.3.3). For the remaining four inference groups, the F1 scores produced by both the classifiers are satisfactory, with values bigger than 60.5%.

8.4.3 Results summarization

When comparing the results of the two scenarios under investigation, we can make the following remarks: the BinaryRelevance and ClassifierChain classifiers give overall low Ham-

Alexia Dini

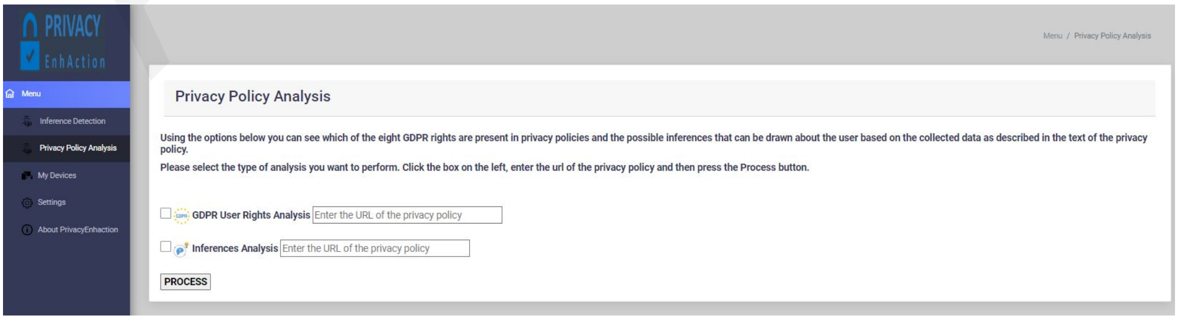
ming Loss scores in both the Rights Classification and Inference Detection components, corresponding to better accuracy of the proposed methods. It appears that the classifiers perform better in terms of precision, recall and F1-measure in the Rights Classification component, as these metrics are higher than those produced by the classifiers in the Inference Detection component, mainly due to the bad performance of the classifiers in the third inference group, as explained above.

In terms of accuracy, the Binary Relevance classifier consistently demonstrates better overall performance, indicating its superiority in accurately classifying the data. The precision, recall, and F1-measure metrics further reinforce the effectiveness of the BinaryRelevance classifier in the Rights Classification component. These metrics exhibit higher values compared to those produced by the Classifier Chain classifier. This suggests that the Binary Relevance classifier achieves a better balance between identifying true positives, minimising false positives, and capturing relevant information related to the rights classification.

In contrast, the performance of the classifiers in the Inference Detection component is relatively weaker, primarily due to the challenges posed by the fourth inference group. However, even in this component, the BinaryRelevance classifier outperforms the ClassifierChain classifier, with slightly lower precision, recall, and F1-measure metrics. This indicates that the Binary Relevance classifier still maintains a superior ability to detect and classify inferences accurately, with some limitations in specific inference groups.

Overall, the results highlight the superiority of the BinaryRelevance classifier over the ClassifierChain classifier in terms of accuracy, precision, recall, and F1-measure metrics. The BinaryRelevance classifier exhibits better performance in both the Rights Classification and Inference Detection components, showcasing its effectiveness in accurately classifying data and enhancing privacy protection in the implemented system.

Alexia Dini



Privacy Policy Analysis

Using the options below you can see which of the eight GDPR rights are present in privacy policies and the possible inferences that can be drawn about the user based on the collected data as described in the text of the privacy policy.

Please select the type of analysis you want to perform. Click the box on the left, enter the url of the privacy policy and then press the Process button.

□ GDPR User Rights Analysis Enter the URL of the privacy policy

□ Inferences Analysis Enter the URL of the privacy policy

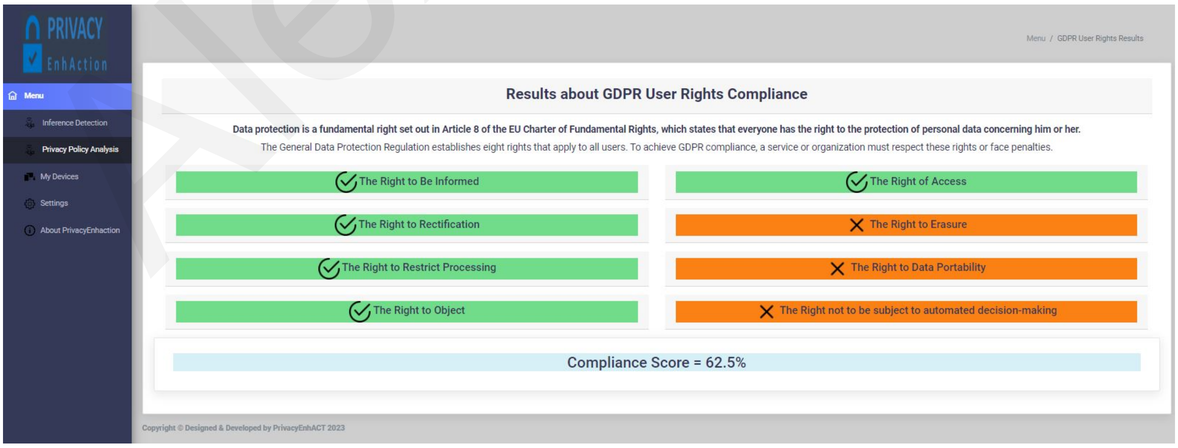
PROCESS

Figure 8.2: PrivacyEnhAction Privacy Policy Analysis page

8.5 Extending PrivacyEnhAction with the SpotAware approach

We have extended the PrivacyEnhAction web application with the functionalities provided by the SpotAware approach using the BinaryRelevance classifier, expanding the options offered to the users, where they can be further informed about which of the eight GDPR user rights are present in privacy policies and the possible inferences that can be drawn about them based on the collected data as described in the text of a privacy policy. Figure 8.2 portrays in a screenshot the new Privacy Policy analysis page of the application, where the users can select either the option for the GDPR user rights analysis, or the inferences analysis.

In order to demonstrate the results of the GDPR User Rights Analysis and the Inferences Analysis components, we have used the privacy policy of the Cubot wearables[2]. In Figure 8.3, one can see the results of the GDPR User Rights Analysis component. The GDPR User Rights that are highlighted in green are those that have been classified as being addressed in the specific privacy policy text, while the User Rights highlighted in orange are those that have not been classified as being addressed by the policy. The total compliance score of the specific privacy policy is also calculated based on the number of GDPR user rights addressed. Given that there are eight user rights, the compliance score is proportional to the number of rights addressed in the policy text, i.e. if 1 user right is addressed, the compliance score is 12.5, if 2 user rights are addressed, the score is 25, and so on.



Results about GDPR User Rights Compliance

Data protection is a fundamental right set out in Article 8 of the EU Charter of Fundamental Rights, which states that everyone has the right to the protection of personal data concerning him or her.

The General Data Protection Regulation establishes eight rights that apply to all users. To achieve GDPR compliance, a service or organization must respect these rights or face penalties.

✓ The Right to Be Informed

✓ The Right of Access

✓ The Right to Rectification

✗ The Right to Erasure

✓ The Right to Restrict Processing

✗ The Right to Data Portability

✓ The Right to Object

✗ The Right not to be subject to automated decision-making

Compliance Score = 62.5%

Copyright © Designed & Developed by PrivacyEnhACT 2023

Figure 8.3: Results of the GDPR User Rights Analysis component from PrivacyEnhAction

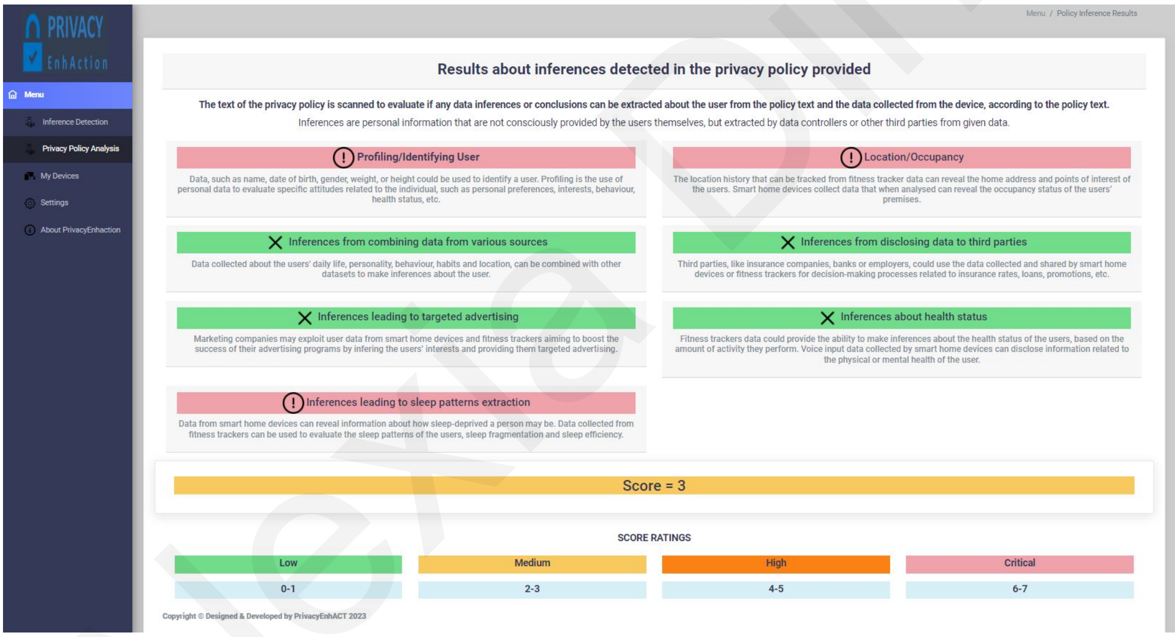
Figure 8.4 depicts the results from the Inferences Analysis component of the PrivacyEnhAction web application for the same privacy policy. Here, the seven inference groups as

Alexia Dini

[2] https://www.cubot.net/platform/About/privacy.html

identified in 8.3.3 can be seen, where the inference groups for which inferences are possible from the privacy policy text are indicated using red colour, while the ones for which no inferences have been detected are marked with the green colour. A score for the privacy policy is also available, which is 3, while the provided score ratings provide explanations regarding how we have evaluated the results. In particular, a score of 0 to 1, means that there is minimal to low risk for inferences from a privacy policy. A score of 2-3 can indicate a Medium risk, 4-5 a High risk and 6-7 a Critical inference risk.

The score for a privacy policy is calculated based on the number of inferences that are possible from the privacy policy text. A score of 0 means that no inferences are possible from the privacy policy text, a score of 1 means that the SpotAware approach for the Inference Detection component that has identified one possible inference, and so on.



Results about inferences detected in the privacy policy provided

The text of the privacy policy is scanned to evaluate if any data inferences or conclusions can be extracted about the user from the policy text and the data collected from the device, according to the policy text.

Inferences are personal information that are not consciously provided by the users themselves, but extracted by data controllers or other third parties from given data.

(!) Profiling/Identifying User

Data, such as name, date of birth, gender, weight, or height could be used to identify a user. Profiling is the use of personal data to evaluate specific attitudes related to the individual, such as personal preferences, interests, behaviour, health status, etc.

(!) Location/Occupancy

The location history that can be tracked from fitness tracker data can reveal the home address and points of interest of the users. Smart home devices collect data that when analysed can reveal the occupancy status of the users' premises.

✕ Inferences from combining data from various sources

Data collected about the users' daily life, personality, behaviour, habits and location, can be combined with other datasets to make inferences about the user.

✕ Inferences from disclosing data to third parties

Third parties, like insurance companies, banks or employers, could use the data collected and shared by smart home devices or fitness trackers for decision-making processes related to insurance rates, loans, promotions, etc.

✕ Inferences leading to targeted advertising

Marketing companies may exploit user data from smart home devices and fitness trackers aiming to boost the success of their advertising programs by infering the users' interests and providing them targeted advertising.

✕ Inferences about health status

Fitness trackers data could provide the ability to make inferences about the health status of the users, based on the amount of activity they perform. Voice input data collected by smart home devices can disclose information related to the physical or mental health of the user.

(!) Inferences leading to sleep patterns extraction

Data from smart home devices can reveal information about how sleep-deprived a person may be. Data collected from fitness trackers can be used to evaluate the sleep patterns of the users, sleep fragmentation and sleep efficiency.

Score = 3

SCORE RATINGS

Low	Medium	High	Critical
0-1	2-3	4-5	6-7

Copyright © Designed & Developed by PrivacyEnhACT 2023

Figure 8.4: Results of the Inferences Analysis component from PrivacyEnhAction

8.6 Discussion

In the validation step of this study, we tested different classifier configurations for our classification tasks in the Rights Classification and Inference Detection components. The results obtained indicate that the proposed method could be useful in the effort to make the users of fitness trackers and smart home devices aware and informed about the data collection practices of their devices in relation to their privacy rights and also educate them about any possible inferences that could be drawn about them from their data. The SpotAware approach

Alexa Dini

is more accurate in detecting specific user rights and inference groups, such as *The Right to Be Informed* or *The Right to Object* GDPR user rights, and the *Location/Occupancy* or the *Inference leading to sleep patterns extraction* inference groups.

A very interesting finding in our study in relation to the evaluation of inferences emerging from privacy policies, is that the fourth inference group we have defined, *Inferences from disclosing data to third parties*, scores very low in the classification process. This has been justified by the difficulty to distinguish between sentences containing text about the rightful disclosure of data to third parties and sentences containing text about the unlawful disclosure of data to other parties, as well as by the lack of sufficient information related to the third party affiliates of the providers. This is in line with previous research about mobile apps, where it is reported that only 22% of them provide the names of their third-party partners in their privacy policy, while 10% provide no information at all [212].

Another noticeable contribution of our research is the identification of the GDPR user rights and inferences that are most commonly found in the privacy policies of fitness trackers and smart home devices. In this study, we created a list of GDPR related terms that should be included in the privacy policy of a fitness tracker or a smart home device, based on the available terms list from the work of Vanezi et al. [300], which we expanded and enriched with more terms after analysing the eight GDPR user rights. The authors in [300] use their terms list to evaluate the GDPR compliance of web platforms from five sectors, with the following compliance scores: (i) Banking=75%, (ii) E-commerce=70.75%, (iii) Education=52.68%, (iv) Travel=72.73% and (v) Social media=69.29%. We also performed a GDPR compliance check for fitness trackers and smart home devices using a similar approach, in order to compare the results, by adopting the list of terms we created, counting the number of times a term categorised under a specific user right appears in each privacy policy file. Figure 8.5 shows the results of the compliance check per sector, while Figure 8.6 shows the average number of terms per GDPR user right that appear in a privacy policy for the two sectors under study, from where we can identify the GDPR user rights that are most commonly found in the privacy policies of fitness trackers and smart home devices.

Compared to the results of the compliance check in [300], we observe that the Smart Home Devices sector has a score of 46.45% and the Fitness Trackers sector a score of 48.77%, which are lower than the compliance scores achieved for the sectors in [300]. The justification for these low scores is twofold. First, we have the inclusion of the eighth GDPR user right, *Right to Avoid Automated Decision-Making* that occurs with a very low average in the privacy policies, i.e. 0.03 terms per fitness trackers privacy policy and 0.24 terms

Alexia Dini

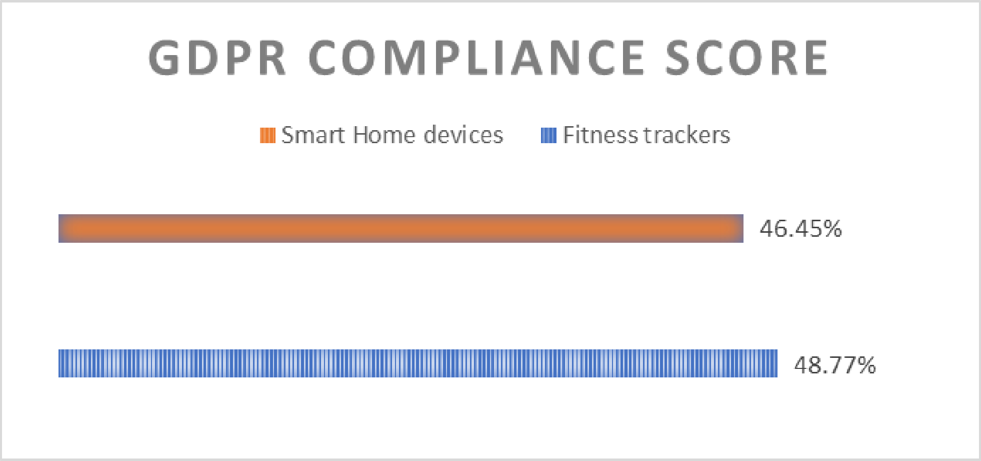


Figure 8.5: Average score of GDPR compliance per sector

per smart home devices privacy policy. Recent research related to the scope of GDPR has showed that in the case of automated decision-making, data controllers do not give details about numerous types of information and in some cases parts of information is seldom or never provided at all [73]. The second reason for the low compliance scores is that there were cases of privacy policies with a 0% compliance score, addressing none of the GDPR user rights.

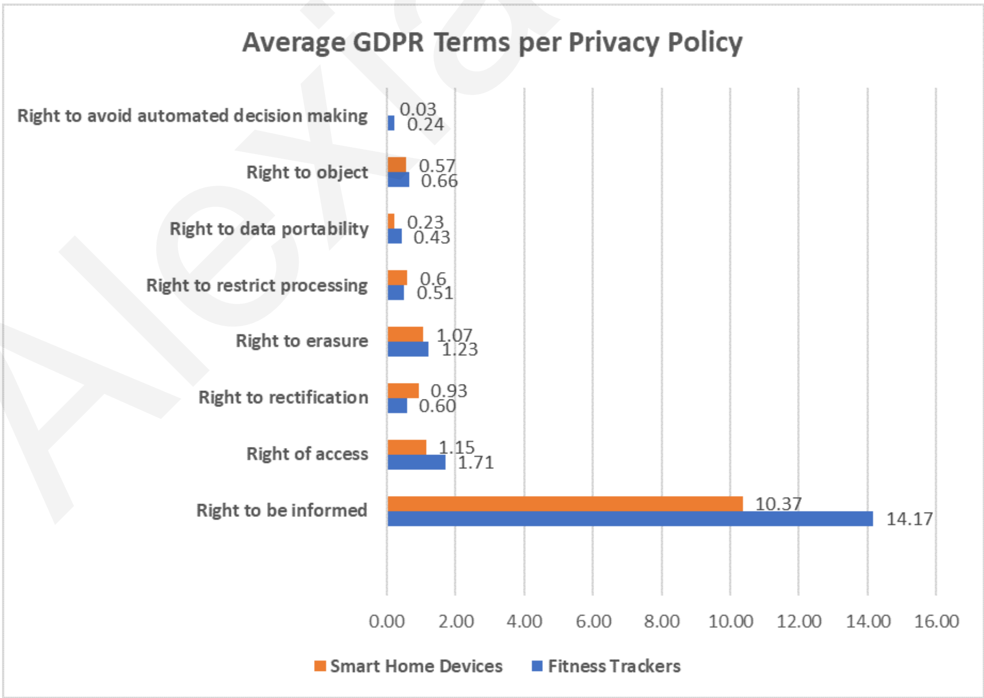


Figure 8.6: Average number of GDPR terms per user right per sector

In relation to the most commonly used GDPR user rights in the privacy policies of fitness trackers and smart home devices, we can see that the *Right to be informed* ranks first with a big difference from the rest, with an average of 14.17 occurrences per privacy policy in the

Alexia Dini

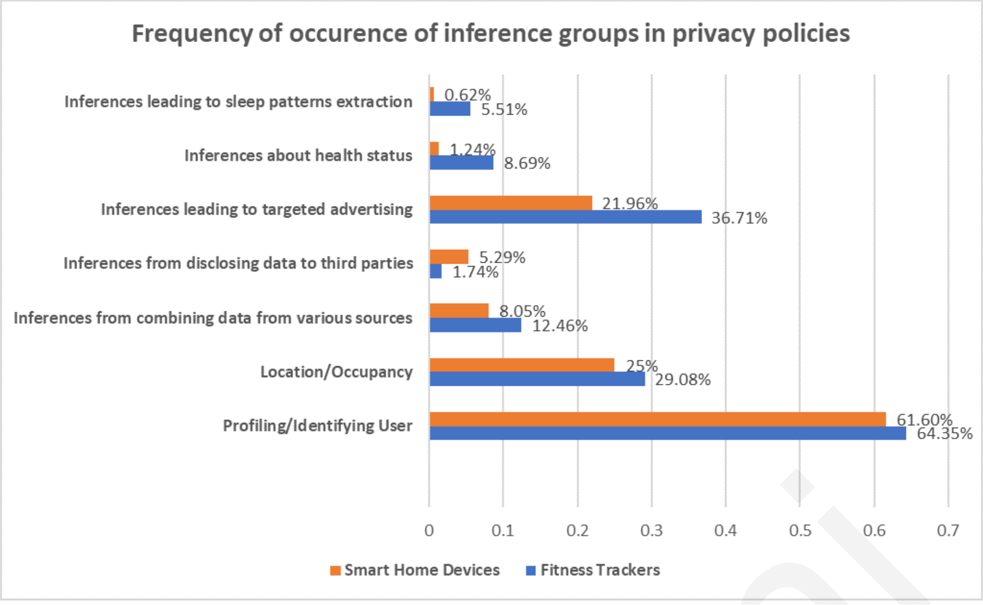


Figure 8.7: Frequency of occurrence of inference groups per privacy policy per sector

fitness trackers domain and 10.37 occurrences in the smart homes domain. The second right is the *Right to access*, with an average of 1.71 occurrences in the fitness trackers domain and 1.15 in the smart homes privacy policies, followed by the *Right to Erasure* with a small deviation.

With reference to the inference groups, by analysing the occurrence of each group in the privacy policies we observe that the most prevalent inference group is *Profiling/Identifying User* that appears in 64.35% of the fitness tracker policies sentences and in 61.60% of the smart home devices sentences, as can be seen in Figure 8.7. This is followed by *Inferences leading to targeted advertising*, with 36.71% for fitness trackers and 21.96% for smart home devices, and by *Location/Occupancy*, appearing in 29.08% of fitness trackers and 25% of smart home devices privacy policy sentences.

Limitations. A challenge that we came across early in the stages of this study is the ambiguity and broadness used in the text of privacy policies, making the labelling process very demanding. The use of misleading, vague language and generic examples in order to legitimise data collection and use practices that are beyond what the user has consented on is quite common and could have a negative effect on the results of our models. A relevant example is the following: "*We may use your personal data for the purposes below: To assess and improve our products and services*". This may be attributed to the fact that automated tools are frequently used in order to produce the policy text, providing such vague expressions to make sure the service provider is legally covered concerning data usage. Regarding

Alexia Dini

external validity, referring to the extent we can generalise our findings, we manually analysed a large dataset of privacy policies creating a corpora of an important size in relevance to previous works, e.g. OPP-115 dataset [317] used in previous works, such as Polisis [125]. Nevertheless, this dataset was not relevant in our case, as it does not contain information on GDPR user rights or data inferences. In terms of other limitations, our study relies widely on the manual labelling process performed and is thus, prone to human error. In order to minimise this bias, a subset of the privacy policies were labelled by more than one experts and all experts had to reach an agreement on the user rights and inference groups present in these policies. Human labelling is nevertheless, a vital part of the classification process and the resulting datasets are one of the contributions of our work.

Alexia Dini

Chapter 9

The implementation of the PrivacyEnhAction application

Alexia Dini

9.1 Introduction

PrivacyEnhAction is a web application that we have developed as part of our PhD research, designed to help users analyse the data collected by smart home devices and fitness trackers in order to identify potential privacy vulnerabilities and inferences that can be drawn from their use. With the growing popularity of these devices, there is an increasing concern about the privacy implications of the data they collect and transmit, and it has become increasingly important to educate users about the data they collect and how they are being shared and used.

PrivacyEnhAction aims to address this concern by providing users with a tool that enhances their awareness of the information that is being collected by their devices and the potential inferences that can be drawn from these data. By increasing users' knowledge and understanding of the privacy risks associated with these devices, PrivacyEnhAction can help them make more informed decisions about their use. Our goal in developing PrivacyEnhAction was to create a user-friendly and accessible tool that can be used by anyone who wants to better understand the privacy implications of their smart home devices and fitness trackers. We believe that our application will be particularly useful for individuals who are concerned about their privacy and want to take a more proactive approach to protecting it.

In this chapter, we will provide a detailed description of the design, architecture, implementation, features, and evaluation of PrivacyEnhAction, highlighting its potential benefits for users and its contribution to the field of privacy enhancement. PrivacyEnhAction is available online at http://privacyenhaction.com:8989, while the code is available as open-source software under the GPL-3.0 license on a GitHub repository[1].

9.2 PrivacyEnhAction Design and Architecture

9.2.1 Overview of the PrivacyEnhAction architecture

The web application developed in this thesis, PrivacyEnhAction, was created using Flask [216], a popular Python web framework for developing web applications that utilises RESTful APIs to provide web services that conform to the REST architectural style. The application's architecture is based on a client-server model, where the server is implemented using Flask and is responsible for handling incoming HTTP requests and providing responses to clients. To ensure efficient handling of requests and maximise performance, we used the Gunicorn web server [119] to deploy the Flask application. One of the key features of the PrivacyEnhAction web application is the integration of machine learning models for the extraction of information based on user input. These models were trained using datasets of historical data and were integrated into the application using Flask's extensibility features.

Based on the above, there are three layers in the PrivacyEnhAction application: (a) An Apache web server, (b) a Gunicorn WSGI web server, and (c) a Flask web application framework. Their relationships can be seen in Figure 9.1.

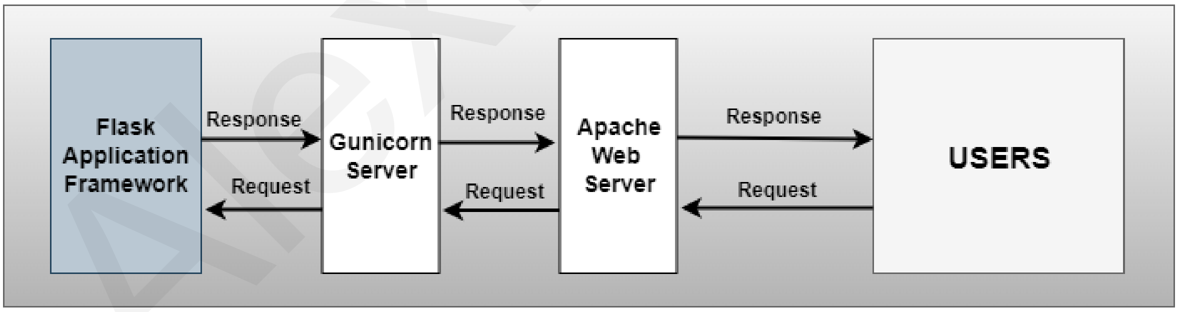


Figure 9.1: PrivacyEnhAction back-end architecture [329]

Alexia Dini

9.2.2 Overview of the PrivacyEnhAction design

The design of PrivacyEnhAction is a critical aspect of the application's development, as it determines how the application will function and how users will interact with it. In this section, we describe the design of PrivacyEnhAction in detail. We begin by discussing the underlying

[1] https://github.com/CS-UCY-SEIT-lab/PrivacyEnhAction

technologies used in the application's development, including Flask, Flask-RESTful, Gunicorn, and Apache. We then describe the data storage mechanism used in PrivacyEnhAction, which relies on file storage. Finally, we discuss the application's data analysis techniques and notification system, which help to identify potential privacy risks associated with smart home devices and fitness trackers. By providing an in-depth understanding of the design of PrivacyEnhAction, this section aims to provide insight into the application's functionality and how it meets the needs of its users. The key features of PrivacyEnhAction are the following:

- **Flask Framework:** PrivacyEnhAction is developed using the Flask web framework, which is lightweight, flexible, suitable for creating small projects and easily scalable. Flask provides a simple and flexible interface for creating web applications in Python, with support for a wide range of web development tasks, including routing, rendering templates, and handling HTTP requests and responses.
- **RESTful API with Flask-RESTful:** PrivacyEnhAction's RESTful API is built using Flask-RESTful. This extension simplifies the creation of RESTful endpoints, allowing developers to focus on the application's core functionality.
- **Gunicorn:** Gunicorn is a Python web server gateway interface (WSGI) HTTP server for UNIX, that is used to serve the Flask application. It provides a simple way to deploy Python web applications, allowing multiple workers to handle requests concurrently, each handling requests in parallel, thus improving the overall throughput of the application.
- **Apache Web Server:** Apache is a widely used web server that is used to serve the application to the end-users. It acts as a reverse proxy for Gunicorn, forwarding requests to Gunicorn for processing.
- **File Storage:** Instead of using a database, PrivacyEnhAction stores its data in files. This simplifies the application's design and makes it easier to deploy and manage. The files are organised in a structured format that allows the application to read and write data as needed.
- **Visual representation tools:** Many visual representation tools are being used in PrivacyEnhAction to display information to the users, such as tool-tips, charts, graphs, info-graphics, and pop-up messages. The user interface of PrivacyEnhAction has been designed to be intuitive and easy to use, with clear labels and visual cues to help users understand the application's functionality, following Nielsen and Molich's 10 user interface design guidelines [200] retaining all graphic representations and text across

Alexia Dini

every system template.

9.3 PrivacyEnhaction Implementation

9.3.1 Back-end development of PrivacyEnhAction

The back-end of our web application was implemented using the Flask web framework and a variety of Python libraries and tools. To start, we used Flask to create the web server that hosts our application. To deploy our Flask-based application, we used Gunicorn, that allowed us to deploy our application with minimal configuration, while also providing support for handling multiple simultaneous requests. All functionality is available to the users without the need for a registered account.

In our web application, we utilised file storage as an alternative to using a traditional database system. This decision was made based on the specific requirements of our application, as well as the advantages that file storage provides over a database system. In our case, we needed to store and access a large number of files, including images, that are dynamically created by the application. While it is possible to store these files in a database, this can often result in slower performance and increased complexity, especially when dealing with large files or a high volume of file uploads. To avoid these issues, we opted to store our files directly on the file-system. This allowed us to easily upload, store, and access files from within our application, without the need for a separate database system. Using file storage also provided us with a number of additional benefits, including easier backup and recovery of our data, simplified data management, and reduced costs compared to using a database system. While file storage may not be suitable for all types of web applications, it can provide a viable alternative to traditional database systems in cases where file-based data storage and retrieval are the primary concerns. By utilising Flask's built-in file handling functionality, we were able to build a web application that met our specific needs and provided a fast, reliable, and scalable solution for managing and accessing our files.

The scripts of the PrivacyEnhAction modules were written in the Python programming language without any external Python library dependencies. In order to train and integrate our machine learning models into the application, we have utilized the Python Scikit-learn library. Scikit-learn is a widely used Python library for the implementation of machine learning models and statistical modelling, providing support for a wide range of machine learning algorithms and techniques, as well as functionalities for dimensionality reduction, feature

Alexia Dini

selection, feature extraction, ensemble techniques, and inbuilt datasets. We have also used Matplotlib, which is a comprehensive library for creating static, animated, and interactive visualisations in Python.

9.3.2 Front-end development of PrivacyEnhAction

The front-end of PrivacyEnhAction was implemented using a combination of HTML, CSS, and JavaScript, with the help of several third-party libraries and tools. To begin, we used HTML to create the basic structure of our web pages, including the layout, text, images, and other content. HTML provides a simple and flexible markup language for creating web pages, with support for a wide range of elements and attributes. Next, we used CSS to style our web pages, including the fonts, colours, backgrounds, borders, and other visual aspects. CSS allowed us to create a visually appealing and consistent design for our application, with support for responsive design and other advanced features. Finally, we used JavaScript to add interactivity and dynamic behaviour to our web pages, including user input validation, and dynamic content loading. JavaScript provides a powerful and flexible scripting language for building complex web applications, with support for a wide range of APIs and libraries.

To simplify the development process and speed up development, we utilised several third-party libraries and tools, including Bootstrap and Flask-Bootstrap. Bootstrap is a popular CSS framework that provides pre-built User Interface components and styles for building responsive web applications. Flask-Bootstrap is a Flask extension that provides integration between Bootstrap and Flask, making it easy to incorporate Bootstrap into our Flask-based application. By utilising these technologies and tools, we were able to create a fast, responsive, and visually appealing web application that provides a seamless user experience.

Alexia Dini

9.4 PrivacyEnhAction Features and Modules

The PrivacyEnhAction application has two main modules, Inference Detection Analysis and Privacy Policy Analysis. In this section, we provide an overview of the features and functionality of each module.

9.4.1 Inference Detection Analysis module

The Inference Detection Analysis module is designed to identify potential privacy risks associated with the data collected by specific smart home devices and fitness trackers. In order to achieve this, the module incorporates the K-Means clustering algorithm, which assists in analysing and identifying patterns within the collected data from smart home devices. The implementation of the K-Means algorithm leverages the functionality described in Chapter 4, where the clustering technique was explored in depth. By applying K-Means clustering within the Inference Detection Analysis module, the web application gains the ability to group data points based on their similarities, allowing for the detection of potential inferences that can be drawn from the collected data. The algorithm groups data points into meaningful clusters, allowing for the identification of potentially sensitive patterns or associations that could compromise user privacy.

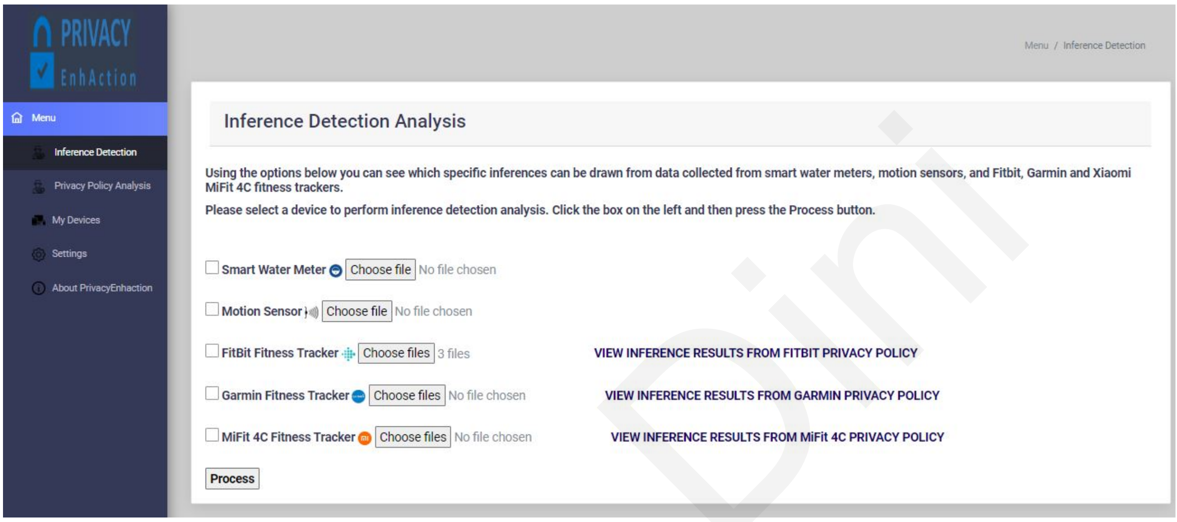
In addition to the K-Means algorithm, the PrivacyEnhAction web application employs a range of data analysis techniques, statistical analysis methods, and descriptive analytics that were described in Chapter 5, to identify the inferences that can be extracted about the users from the data collected by their fitness trackers. These techniques play a vital role in extracting valuable insights and uncovering potential privacy risks. By leveraging these analytics, the application can identify common patterns, behavioural tendencies, and potential privacy-invasive inferences. By combining data analysis, statistical analysis, and descriptive analytics techniques, the PrivacyEnhancement web application can effectively identify inferences that can be drawn about users from the data collected by their fitness trackers. These techniques enable the detection of sensitive information, such as daily routines, activity levels, sleep patterns, or health conditions, which may pose potential privacy risks if disclosed without user consent. Through the use of these techniques, the PrivacyEnhAction web application empowers users to gain insights into the potential privacy risks associated with their fitness tracker data. Users can better understand the inferences that can be derived from their data, enabling them to make informed decisions about data sharing, privacy settings, and overall data protection.

The module then informs the user about the potential privacy risks based on these inferences. The user can view the analysis results in a dedicated dashboard in the form of text, charts, graphs, tool-tips, info-graphics, and pop-up messages, that inform the user about the possible inferences and privacy risks, depending on the device. Furthermore, the dashboard presents the users with educational information about the relevant inferences and privacy risks and supplements this educational effort with the provision of relevant links that the user can select to read for further information. The educational information presented to the users is based on various sources, as for example health related portals, like the *Sleep Foundation* [265], a source for evidence-based, medically reviewed sleep health information and

Alexia Dini

in-depth product testing, or *Healthline* [127], which is about *“making health and wellness information accessible, understandable, and actionable so that readers can make the best possible decisions about their health"*.

Figure 9.2 shows the initial page of the PrivacyEnhAction Inference Detection Analysis module. The user can select one of the available devices by ticking the box on the left and then by pressing the “*Choose files*" button, she can upload the appropriate files for analysis.



Inference Detection Analysis

Using the options below you can see which specific inferences can be drawn from data collected from smart water meters, motion sensors, and Fitbit, Garmin and Xiaomi MiFit 4C fitness trackers.

Please select a device to perform inference detection analysis. Click the box on the left and then press the Process button.

□ Smart Water Meter [Choose file] No file chosen

□ Motion Sensor [Choose file] No file chosen

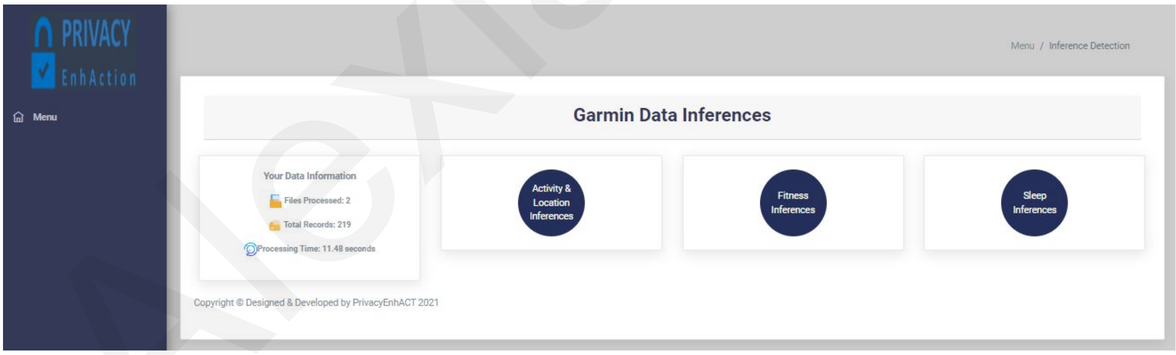
□ FitBit Fitness Tracker [Choose files] 3 files VIEW INFERENCE RESULTS FROM FITBIT PRIVACY POLICY

□ Garmin Fitness Tracker [Choose files] No file chosen VIEW INFERENCE RESULTS FROM GARMIN PRIVACY POLICY

□ MiFit 4C Fitness Tracker [Choose files] No file chosen VIEW INFERENCE RESULTS FROM MiFit 4C PRIVACY POLICY

[Process]

Figure 9.2: PrivacyEnhAction Inference Detection Analysis page



Garmin Data Inferences

Your Data Information

Files Processed: 2

Total Records: 219

Processing Time: 11.48 seconds

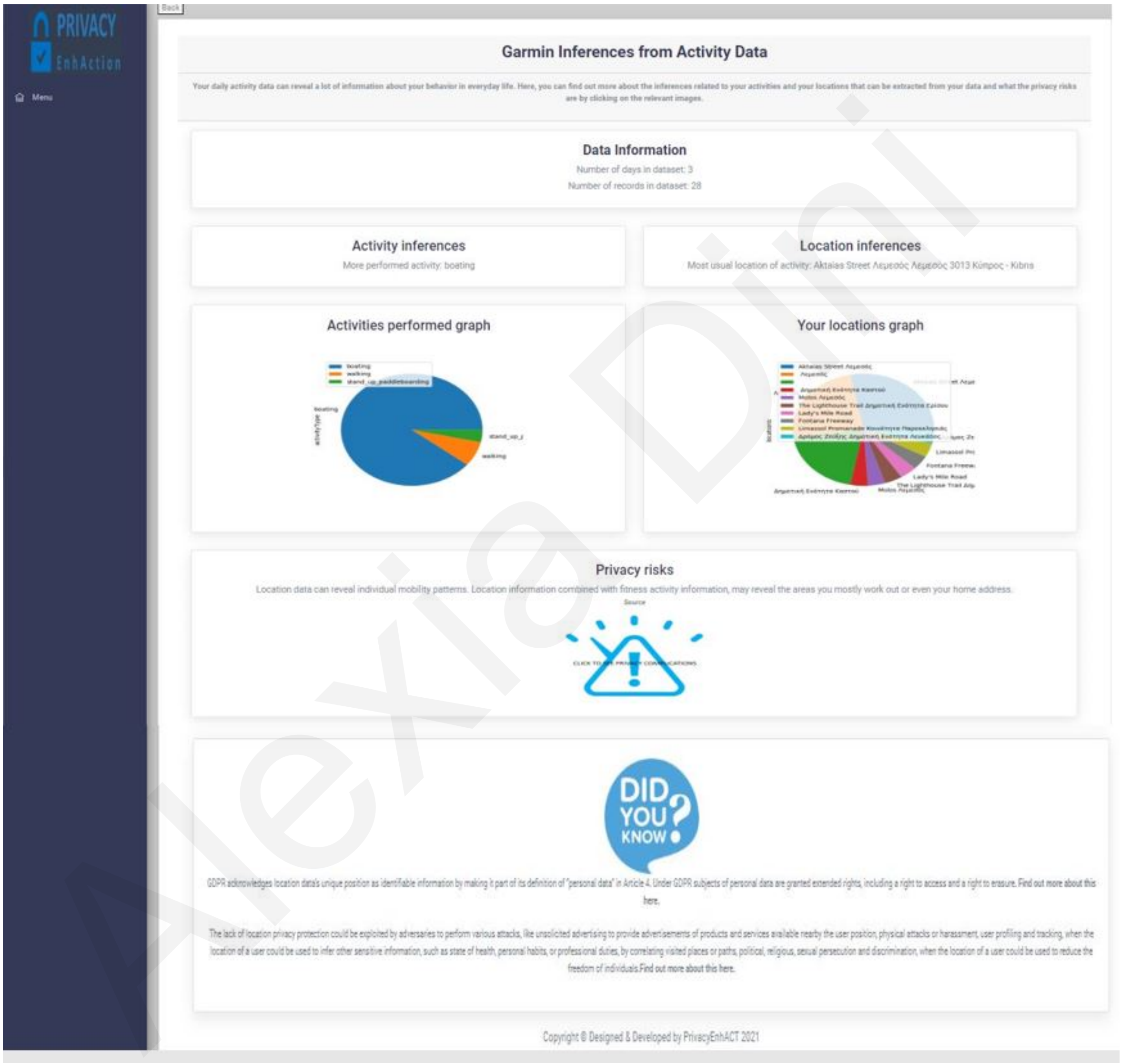
Copyright © Designed & Developed by PrivacyEnhACT 2021

Figure 9.3: Garmin Inference Detection Analysis results

Alexia Dini

In the screenshot provided in Figure 9.3, we can see the inference detection analysis results page of the Garmin dataset we have created in Chapter 5. As already mentioned in Chapter 5, Section 5.3.2, the inference detection analysis process performed on the Garmin dataset provides activity and location related inferences, fitness related inferences and sleep related inferences, as can be seen in Figure 9.3. The user can press the corresponding button to view the relevant dashboard.

Figure 9.4 shows the dashboard with the results for the Garmin activity and location related inferences, while Figure 9.5 shows the dashboard with the results for the fitness



Garmin Inferences from Activity Data

Your daily activity data can reveal a lot of information about your behavior in everyday life. Here, you can find out more about the inferences related to your activities and your locations that can be extracted from your data and what the privacy risks are by clicking on the relevant images.

Data Information

Number of days in dataset: 3

Number of records in dataset: 28

Activity inferences

More performed activity: boating

Location inferences

Most usual location of activity: Aktaias Street Λεμεσός Λεμεσός 3013 Κύπρος - Kıbrıs

Activities performed graph

Your locations graph

Privacy risks

Location data can reveal individual mobility patterns. Location information combined with fitness activity information, may reveal the areas you mostly work out or even your home address.

Source

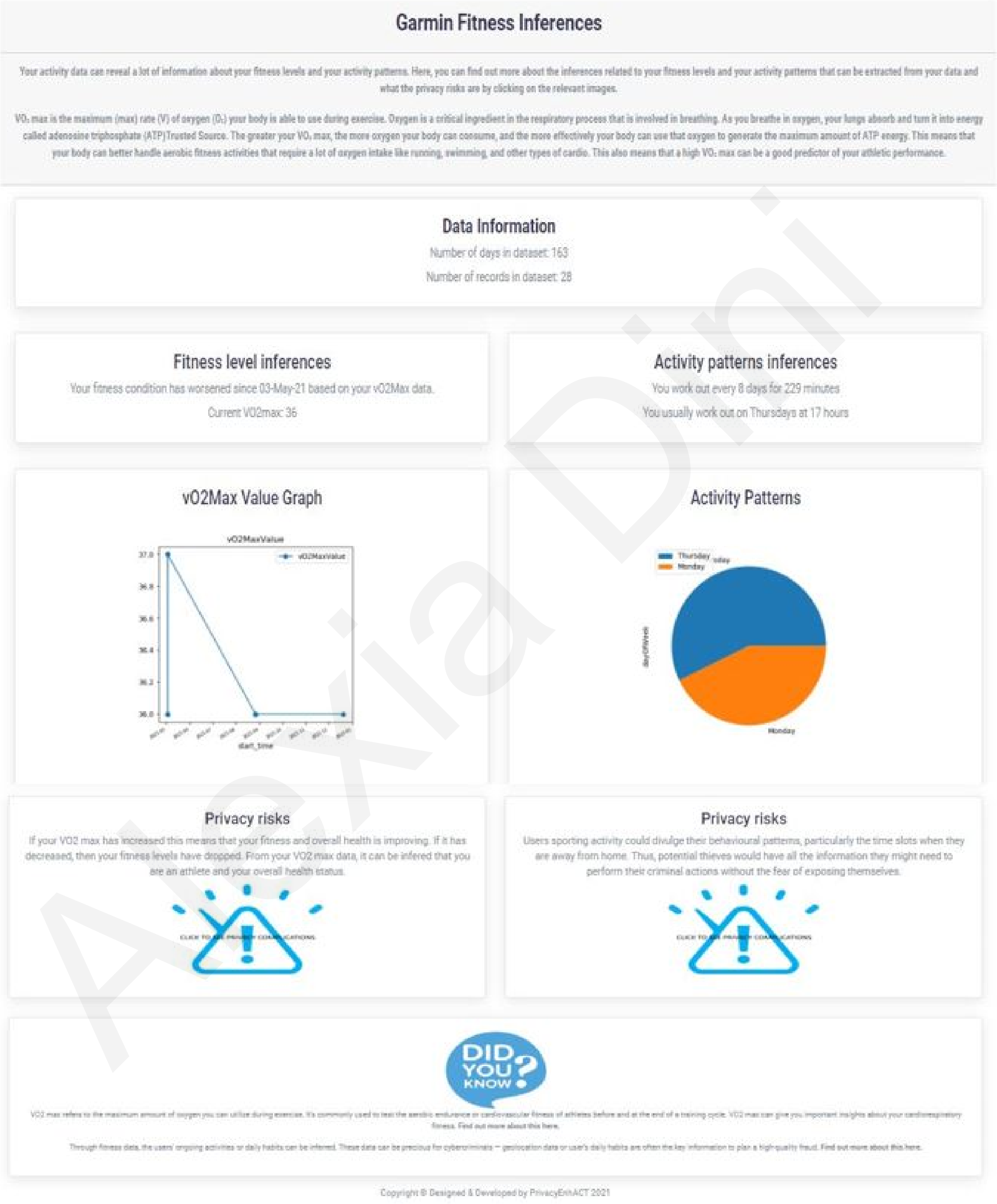
GDPR acknowledges location data's unique position as identifiable information by making it part of its definition of "personal data" in Article 4. Under GDPR subjects of personal data are granted extended rights, including a right to access and a right to erasure. Find out more about this here.

The lack of location privacy protection could be exploited by adversaries to perform various attacks, like unsolicited advertising to provide advertisements of products and services available nearby the user position, physical attacks or harassment, user profiling and tracking, when the location of a user could be used to infer other sensitive information, such as state of health, personal habits, or professional duties, by correlating visited places or paths, political, religious, sexual persecution and discrimination, when the location of a user could be used to reduce the freedom of individuals.Find out more about this here.

Copyright © Designed & Developed by PrivacyEnhACT 2021

Figure 9.4: Garmin Activity and Location Inferences

Alexia Dini



Garmin Fitness Inferences

Your activity data can reveal a lot of information about your fitness levels and your activity patterns. Here, you can find out more about the inferences related to your fitness levels and your activity patterns that can be extracted from your data and what the privacy risks are by clicking on the relevant images.

VO₂ max is the maximum (max) rate (V) of oxygen (O₂) your body is able to use during exercise. Oxygen is a critical ingredient in the respiratory process that is involved in breathing. As you breathe in oxygen, your lungs absorb and turn it into energy called adenosine triphosphate (ATP)Trusted Source. The greater your VO₂ max, the more oxygen your body can consume, and the more effectively your body can use that oxygen to generate the maximum amount of ATP energy. This means that your body can better handle aerobic fitness activities that require a lot of oxygen intake like running, swimming, and other types of cardio. This also means that a high VO₂ max can be a good predictor of your athletic performance.

Data Information

Number of days in dataset: 163

Number of records in dataset: 28

Fitness level inferences

Your fitness condition has worsened since 03-May-21 based on your vO2Max data.

Current VO2max: 36

Activity patterns inferences

You work out every 8 days for 229 minutes

You usually work out on Thursdays at 17 hours

vO2Max Value Graph

Activity Patterns

Privacy risks

If your VO2 max has increased this means that your fitness and overall health is improving. If it has decreased, then your fitness levels have dropped. From your VO2 max data, it can be infered that you are an athlete and your overall health status.

Privacy risks

Users sporting activity could divulge their behavioural patterns, particularly the time slots when they are away from home. Thus, potential thieves would have all the information they might need to perform their criminal actions without the fear of exposing themselves.

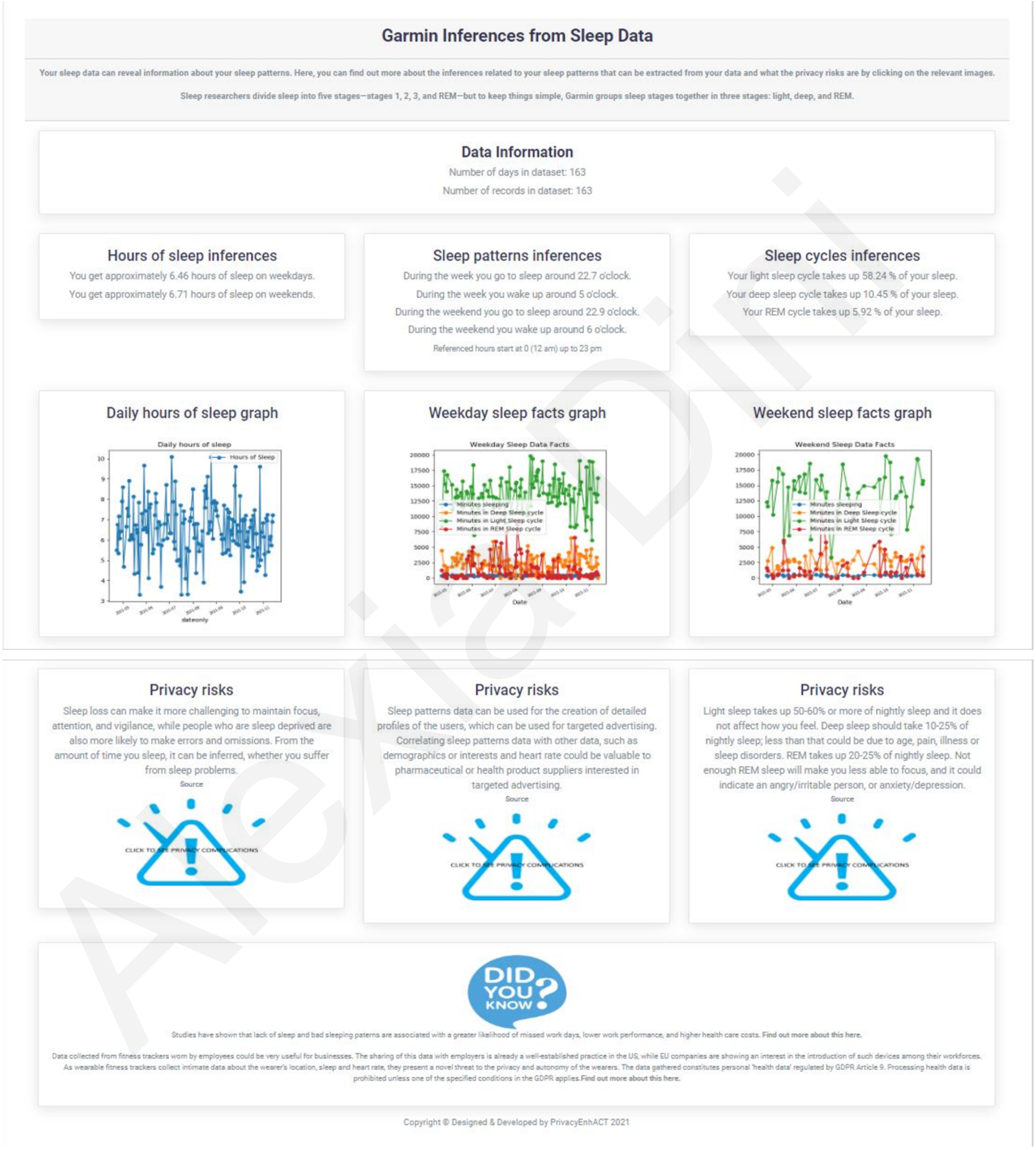
VO2 max refers to the maximum amount of oxygen you can utilize during exercise. It's commonly used to test the aerobic endurance or cardiovascular fitness of athletes before and at the end of a training cycle. VO2 max can give you important insights about your cardiorespiratory fitness. Find out more about this here.

Through fitness data, the users' ongoing activities or daily habits can be inferred. These data can be precious for cybercriminals — geolocation data or user's daily habits are often the key information to plan a high-quality fraud. Find out more about this here.

Copyright © Designed & Developed by PrivacyEnhACT 2021

Figure 9.5: Garmin Fitness Inferences

Alexia Dini



Garmin Inferences from Sleep Data

Your sleep data can reveal information about your sleep patterns. Here, you can find out more about the inferences related to your sleep patterns that can be extracted from your data and what the privacy risks are by clicking on the relevant images.

Sleep researchers divide sleep into five stages—stages 1, 2, 3, and REM—but to keep things simple, Garmin groups sleep stages together in three stages: light, deep, and REM.

Data Information

Number of days in dataset: 163

Number of records in dataset: 163

Hours of sleep inferences

You get approximately 6.46 hours of sleep on weekdays.

You get approximately 6.71 hours of sleep on weekends.

Sleep patterns inferences

During the week you go to sleep around 22.7 o'clock.

During the week you wake up around 5 o'clock.

During the weekend you go to sleep around 22.9 o'clock.

During the weekend you wake up around 6 o'clock.

Referenced hours start at 0 (12 am) up to 23 pm

Sleep cycles inferences

Your light sleep cycle takes up 58.24 % of your sleep.

Your deep sleep cycle takes up 10.45 % of your sleep.

Your REM cycle takes up 5.92 % of your sleep.

Daily hours of sleep graph

Weekday sleep facts graph

Weekend sleep facts graph

Privacy risks

Sleep loss can make it more challenging to maintain focus, attention, and vigilance, while people who are sleep deprived are also more likely to make errors and omissions. From the amount of time you sleep, it can be inferred, whether you suffer from sleep problems.

Source

Privacy risks

Sleep patterns data can be used for the creation of detailed profiles of the users, which can be used for targeted advertising. Correlating sleep patterns data with other data, such as demographics or interests and heart rate could be valuable to pharmaceutical or health product suppliers interested in targeted advertising.

Source

Privacy risks

Light sleep takes up 50-60% or more of nightly sleep and it does not affect how you feel. Deep sleep should take 10-25% of nightly sleep; less than that could be due to age, pain, illness or sleep disorders. REM takes up 20-25% of nightly sleep. Not enough REM sleep will make you less able to focus, and it could indicate an angry/irritable person, or anxiety/depression.

Source

Studies have shown that lack of sleep and bad sleeping paterns are associated with a greater likelihood of missed work days, lower work performance, and higher health care costs. **Find out more about this here.**

Data collected from fitness trackers worn by employees could be very useful for businesses. The sharing of this data with employers is already a well-established practice in the US, while EU companies are showing an interest in the introduction of such devices among their workforces. As wearable fitness trackers collect intimate data about the wearer's location, sleep and heart rate, they present a novel threat to the privacy and autonomy of the wearers. The data gathered constitutes personal 'health data' regulated by GDPR Article 9. Processing health data is prohibited unless one of the specified conditions in the GDPR applies.**Find out more about this here.**

Copyright © Designed & Developed by PrivacyEnhACT 2021

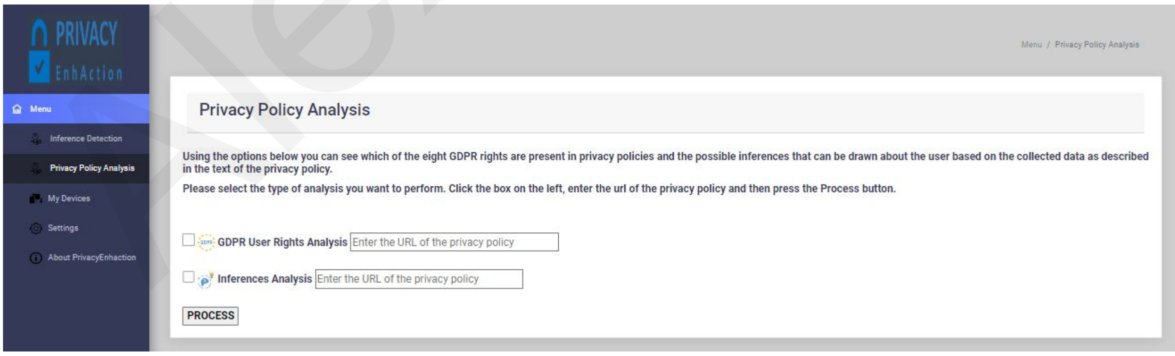
Figure 9.6: Garmin Sleep Inferences

related inferences. Lastly, in Figure 9.6, the sleep related inferences that can be extracted about the user from the Garmin fitness tracker data can be seen.

In all three dashboards, the user is presented with textual information, tool-tips, graphs, charts and info-graphics regarding the extracted inferences, as well as further educative information aiming to provide additional knowledge to the user.

9.4.2 Privacy Policy Analysis module

The Privacy Policy Analysis module has a double purpose. It is designed to analyse the privacy policies of smart home devices and fitness trackers, with the two aims: first to inform the users about which of the eight GDPR user rights are addressed in the privacy policy text, and second, to inform the users about the possible inferences that can be drawn about them based on the collected data as described in the text of a privacy policy. The module incorporates the BinaryRelevance classifier that was described in Chapter 8 to achieve its objectives. By incorporating the Binary Relevance classifier within the Privacy Policy Analysis module, the application empowers users to make informed decisions regarding the use of smart home devices and fitness trackers. It provides them with valuable information about the alignment of the privacy policies of these devices with the GDPR user rights and the potential privacy risks associated with the data collected. This allows users to assess the level of transparency and control offered by the smart home devices or fitness trackers and make privacy-conscious choices in line with their preferences and concerns.



Privacy Policy Analysis

Using the options below you can see which of the eight GDPR rights are present in privacy policies and the possible inferences that can be drawn about the user based on the collected data as described in the text of the privacy policy.

Please select the type of analysis you want to perform. Click the box on the left, enter the url of the privacy policy and then press the Process button.

□ GDPR User Rights Analysis Enter the URL of the privacy policy

□ Inferences Analysis Enter the URL of the privacy policy

PROCESS

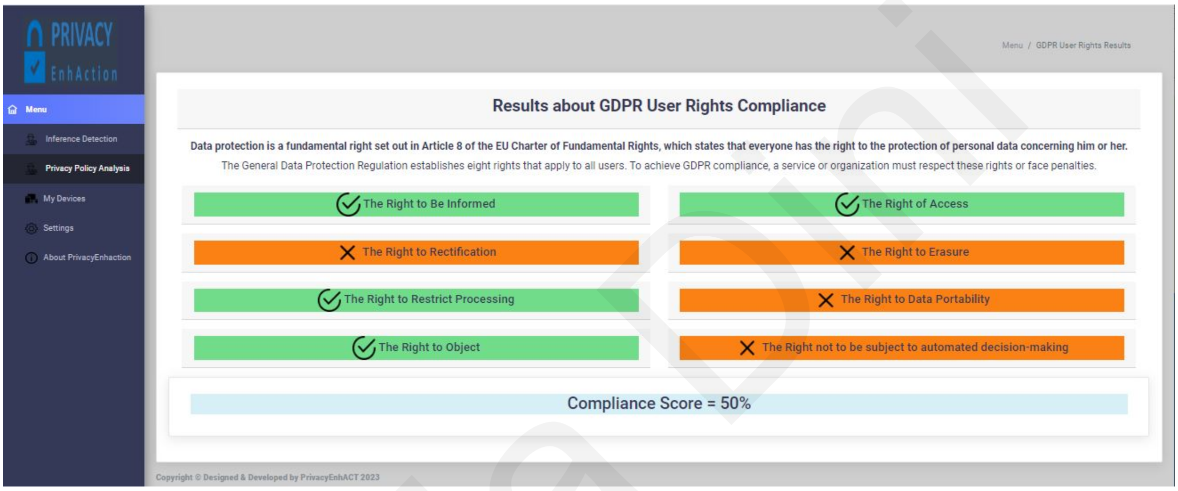
Figure 9.7: PrivacyEnhAction Privacy Policy Analysis page

This module also uses the *BeautifulSoup* and *sentence-splitter* Python libraries and natural language processing techniques for the analysis of the privacy policies texts and to extract the required information from the privacy policies. As already described in Chapter 8, we have trained a combined fitness tracker and smart home devices dataset for the Rights Classification component and the Inference Detection component of the SpotAware approach

Alexia Dini

which we have incorporated in the PrivacyEnhAction application, using two pickle files, created from the two classification trained models, that are used for reloading our trained machine learning models.

In Figure 9.7, we can see the page of the PrivacyEnhAction Privacy Policy Analysis module. Here, the users can select which of the two analyses they wish to perform. For the GDPR User Rights Analysis, the user has to tick the box left to this option, enter the URL of the privacy policy of the fitness tracker or smart home device for which she wants to run the analysis and press the Process button.



Results about GDPR User Rights Compliance

Data protection is a fundamental right set out in Article 8 of the EU Charter of Fundamental Rights, which states that everyone has the right to the protection of personal data concerning him or her.

The General Data Protection Regulation establishes eight rights that apply to all users. To achieve GDPR compliance, a service or organization must respect these rights or face penalties.

✓ The Right to Be Informed	✓ The Right of Access
✗ The Right to Rectification	✗ The Right to Erasure
✓ The Right to Restrict Processing	✗ The Right to Data Portability
✓ The Right to Object	✗ The Right not to be subject to automated decision-making

Compliance Score = 50%

Copyright © Designed & Developed by PrivacyEnhACT 2023

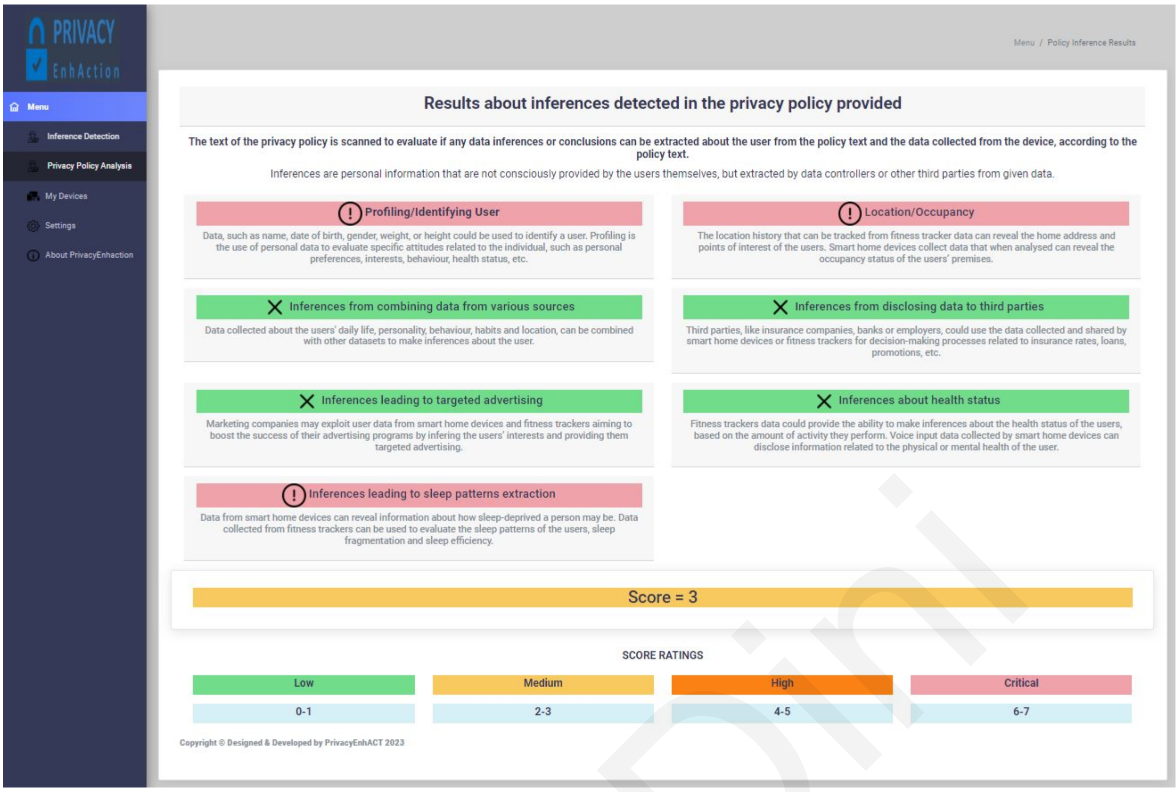
Figure 9.8: *Results about GDPR User Rights Compliance for Withings*

Alexia Dini

In Figure 9.8, the user has chosen to run the GDPR User Rights Analysis of the Withings privacy policy[2], which applies to all users of the Withings App, that is used to connect to the company's wearables and smart home devices, like scales. The GDPR user rights highlighted in the green colour are the rights that are being addressed in this privacy policy, while the ones in orange are not. The overall Compliance Score of the privacy policy in terms of how many GDPR user rights it address is also displays, and in this particular case, the compliance score is 50%, as 4 out of the 8 GDPR user rights are addressed in the policy text.

Next, in Figure 9.9, we see the results of the analysis when the user has chosen to run the privacy policy analysis for the inference detection. In this dashboard, the user can be informed about the possible inferences that can be extracted about them that are highlighted with the red colour, while the ones in green could not be identified from the privacy policy text. The overall score provided for the specific privacy policy is 3, which is classified as Medium, as can be explained at the bottom layer of the dashboard.

[2] https://www.withings.com/us/en/legal/privacy-policy



Results about inferences detected in the privacy policy provided

The text of the privacy policy is scanned to evaluate if any data inferences or conclusions can be extracted about the user from the policy text and the data collected from the device, according to the policy text.

Inferences are personal information that are not consciously provided by the users themselves, but extracted by data controllers or other third parties from given data.

! Profiling/Identifying User

Data, such as name, date of birth, gender, weight, or height could be used to identify a user. Profiling is the use of personal data to evaluate specific attitudes related to the individual, such as personal preferences, interests, behaviour, health status, etc.

! Location/Occupancy

The location history that can be tracked from fitness tracker data can reveal the home address and points of interest of the users. Smart home devices collect data that when analysed can reveal the occupancy status of the users' premises.

✗ Inferences from combining data from various sources

Data collected about the users' daily life, personality, behaviour, habits and location, can be combined with other datasets to make inferences about the user.

✗ Inferences from disclosing data to third parties

Third parties, like insurance companies, banks or employers, could use the data collected and shared by smart home devices or fitness trackers for decision-making processes related to insurance rates, loans, promotions, etc.

✗ Inferences leading to targeted advertising

Marketing companies may exploit user data from smart home devices and fitness trackers aiming to boost the success of their advertising programs by infering the users' interests and providing them targeted advertising.

✗ Inferences about health status

Fitness trackers data could provide the ability to make inferences about the health status of the users, based on the amount of activity they perform. Voice input data collected by smart home devices can disclose information related to the physical or mental health of the user.

! Inferences leading to sleep patterns extraction

Data from smart home devices can reveal information about how sleep-deprived a person may be. Data collected from fitness trackers can be used to evaluate the sleep patterns of the users, sleep fragmentation and sleep efficiency.

Score = 3

SCORE RATINGS

Low	Medium	High	Critical
0-1	2-3	4-5	6-7

Copyright © Designed & Developed by PrivacyEnhACT 2023

Figure 9.9: Results about Inference Detection for Withings

9.5 Further Work

While the current version of PrivacyEnhAction provides valuable tools to increase user awareness about the potential privacy risks associated with smart home devices and fitness trackers, there is still more work to be done to improve the application's functionality and usability. The following are some of the areas where additional work is needed:

- **User Authorisation:** Currently, PrivacyEnhAction does not have a user authorisation system in place. Implementing user authorisation would allow users to create accounts, store their analysis results, and set up their fitness trackers and smart home devices within the application. This would also enable users to keep their data and analysis private and secure.
- **Settings Section**: The addition of a settings section in the application would allow users to customise their analysis settings according to their needs. For example, users could choose the frequency of data analysis or specify which types of data should be analysed. This would provide users with more control over their privacy preferences and enhance the overall user experience.
- **Notification System:** A notification system could be implemented to alert users of any privacy risks detected by the application. This would enable users to take action

Alexia Dini

to protect their privacy and ensure that they are aware of any potential risks in a timely manner.

- **Data Interoperability:** Currently, PrivacyEnhAction only supports a limited number of fitness trackers and smart home devices. Adding support for additional devices would increase the application's interoperability and provide users with a wider range of options. This would also increase the application's potential user base.

Overall, the implementation of these additional features would significantly enhance the functionality and usability of the PrivacyEnhAction application, providing users with a more comprehensive tool to increase their awareness of potential privacy risks associated with smart home devices and fitness trackers.

9.6 Conclusion

The PrivacyEnhAction web application was designed and developed to address the potential privacy vulnerabilities that arise from the use of smart home devices and fitness trackers. The application provides users with valuable tools to increase their awareness about the data collected and shared by their devices, as well as the possible inferences that can be extracted from these data. The design and architecture of the application were carefully considered to ensure that the application was efficient, scalable, and user-friendly. The use of Flask, RESTful API, Gunicorn, and Apache provided a robust framework for building the application, while the incorporation of visual representation tools such as tool-tips, charts, and graphs enhance the user experience.

The two main modules of the application, Inference Detection Analysis and Privacy Policy Analysis, provide users with the ability to analyse their data and identify potential privacy risks. The results of these analyses are presented in a clear and concise manner, informing the users about the privacy risks of using fitness trackers and smart home devices, empowering them to make informed decisions about their privacy, which is one of the most important requirements of the GDPR.

While the current version of the application provides valuable functionality, there is still more work to be done to enhance the user experience and improve the application's overall functionality. The addition of features such as user authorisation, a settings section, a notification system, and increased data interoperability would significantly improve the application's functionality and usability.

In summary, the PrivacyEnhAction web application is a valuable tool for increasing user

Alexia Dini

awareness of potential privacy risks associated with smart home devices and fitness trackers. The application's design, architecture, and functionality were carefully considered and implemented to provide users with a comprehensive tool for analysing their data and making informed decisions about their privacy. With additional work, the application has the potential to become an even more powerful tool for enhancing user privacy in the age of the Internet of Things.

Alexia Dini

Chapter 10

Conclusions and Future Work

Alexia Dini

10.1 Introduction

At the beginning of this thesis, we introduced the challenges associated with the protection of the user data and privacy in the Internet of Things era under the scope of the GDPR: the awareness of the users about how their data are collected and shared by their devices and the privacy risks that occur when the exploitation of such data can lead to the extraction of further information about the users, known as inferences. The research we conducted throughout this thesis has provided valuable insights about the main research problem that we defined in Chapter 1:

Research Problem: How can we increase the smart home devices and fitness trackers users' awareness over their data privacy protection?

Our research has also assisted in answering the 4 research questions that unfolded in the progress:

- **RQ1:** What are the characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess?
- **RQ2:** What inferences can be made from data collected from smart home devices and fitness trackers?
- **RQ3:** Are the users aware of the inferences that can be made about them from their fitness trackers data?
- **RQ4:** Can we enhance the awareness of the users regarding the possible inferences that can be obtained from their fitness tracker data?

In the next sections, we review the goals and findings of this thesis, before moving on to discuss its contributions and to highlight directions for future work.

10.2 Review of Research Goals and Findings

In this thesis we aim to investigate approaches to enhance the users' awareness and control of their data privacy in the GDPR era. Based on the literature review in Chapter 3, we devised the first research question, RQ1, which guided our efforts and research approaches and assisted in the definition of the other three research questions. The knowledge and the insights we derived in relation to each of these research questions are presented in the subsequent chapters as follows: RQ2 in Chapters 4 and 5, RQ3 and RQ4 in Chapter 6. Analysing the insights gained from the research we performed in our effort to address research questions RQ3 and RQ4, it was observed that a big percentage of users of smart home devices and fitness trackers does not read the privacy policies of their devices, adding to the problem of the lack of user awareness of how their data are collected and shared. In Chapter 7 we take one step back in order to investigate what the privacy policies of fitness trackers and smart home devices mention in their text in relevance to the data collection and sharing practices of the companies. Therefore, this chapter presents the analysis we performed on the privacy policies of a number of fitness trackers and smart home devices, aiming to provide clarity and transparency about the privacy practices of these devices, ultimately empowering users to make informed decisions about their data and privacy. Based on the findings obtained from this research, and specifically the problem of the lack of proper communication from the side of service providers to the users in relation to data collection, processing and sharing practices in the privacy policies text, Chapter 8 introduces an approach for the analysis of the privacy policies of such devices in order for the users to be informed about the data collection, processing and sharing practices of the service provider that collects their data, as well as the potential risks that are present from the possible inferences. Finally, Chapter 9 presents the implementation process of the "PrivacyEnhAction" web application that has been developed to support our research efforts towards increasing the user awareness of potential inferences and privacy risks associated with smart home devices and fitness trackers.

10.2.1 What are the characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess?

The answers to RQ1 are presented in Chapter 3. In this chapter we have performed a state-of-the-art literature review regarding the protection of the user privacy in IoT under the GDPR scope, which led to the identification of the need for a user-centric privacy framework for

Alexia Dini

the protection of personal data in the IoT that empowers users with greater control over their personal data. Using four privacy challenges that have been determined by Wachter [308] in how the GDPR principles can safeguard the user privacy in IoT, we have analysed the state-of-the-art literature to define the characteristics that such a framework should own. Each characteristic is based upon the GDPR principles and aims to address the four challenges of data privacy in GDPR, which are “*Profiling, inference and discrimination*", “*Control and context-sensitive sharing of identity*, “*Consent and uncertainty*" and “*Honesty, trust, and transparency*".

The analysis we performed highlights the need to prevent inferences in the processing of personal data, which can reveal more information about a person, such as ethnic origin, political opinions, or even the possibility to uniquely identify a person, which is prohibited in GDPR, according to Article 19. The need to provide data transformation is another issue that we have identified trough our research, as failure in data protection using anonymisation techniques can lead to user tracking or enabling the linking to other data sets. This issue is recognised in GDPR, where data transformation techniques are required to protect user privacy. Under the GDPR, transparency is a fundamental principle that governs the processing of personal data. The GDPR places a high value on individuals being informed about how their personal data is being used, by whom, and for what purpose, therefore we have identified the need for providing increased user awareness on data collection as well as control over their personal data and their devices that collect data, through the research we performed. As the GDPR is all about the users, through our research presented in Chapter 3, we have recognised the urgency for the development of appropriate tools that allow the users to control the usage of their data, oversee and control how they generate and share data, and provide transparency.

The right to erasure, also known as “*The right to be forgotten*", is one of the data subject rights granted under the GDPR, which gives individuals the right to request that their personal data be deleted or erased by the data controller in certain circumstances. Our research emphasises the need for the users to be able to exercise this right, along with the key GDPR requirement for transparency, where users should be given control and be aware of how their data are processed, as transparency is vital for increasing users trust in an IoT system. In reporting these findings, we have to state that from the literature review performed it became obvious that there is a mismatch on how the interests of the user are balanced against those of the third party involved, an issue that has to be addressed in a privacy framework for IoT. Furthermore, users must be able to enforce their privacy preferences in such an environment,

Alexia Dini

where additionally privacy by default or by design must be provided.

During our research, we observed that informed consent is also a a fundamental principle of GDPR that must be addressed, as it governs the processing of personal data, requiring that individuals provide explicit and informed consent before their personal data are collected, processed, or shared. However, in order for the users to be able to provide their informed consent, they must be aware of the privacy risks of data collection and the possible inferences that can be extracted about them from the collected data. This requirement has been the motivation for the rest of the research in this thesis and the definition of the research questions that followed.

10.2.2 What inferences can be made from data collected from smart home devices and fitness trackers?

With RQ2, the aim is to explore the possible inferences that can be extracted about the users from the data collected by their smart home devices and fitness trackers, informing the research community with the results, as found in Chapter 4 and Chapter 5.

To address RQ2, Chapter 4 presents a smart home scenario as the experimental setup where the concept of inference detection is explored, in order to investigate whether the exploitation of data generated by smart water meters and motion sensors can lead to inferences about the occupants routines or other sensitive information. By reviewing the related literature, we identified a number of possible inferences that can be extracted from the use of these smart devices, and through the proposed methodology we aimed to extract inferences in relation to the specific household's occupants routines, based on the collected data, and in particular the time in the morning that the residents wake up, the usual time that they go to sleep at night, to get insights as to whether they wake-up during the night, the time they leave for work in the morning and the time they return.

The use of machine learning methods assisted in the extraction and verification of these inferences with the household residents. Interestingly, one of the experiments conducted in this research involved combining the data from both motion sensors and water consumption sensors in order to derive additional insights. By linking these two sets of data, we were able to identify times when motion and water consumption occurred concurrently or were non-existent. This combination of data allowed for more nuanced inferences to be made about individuals' behaviour and habits. For example, we were able to identify instances where the user woke up during the night to use the toilet, which could potentially indicate

Alexia Dini

underlying health problems if it occurred frequently. Overall, the integration of data from multiple sources allowed for more sophisticated analyses and insights to be derived. This approach highlights the potential for further research to combine data from a variety of smart home devices and fitness trackers in order to gain a more comprehensive understanding of the potential inferences that can be extracted about the individuals' behaviour, health, etc. By leveraging the power of machine learning and data analytics, we see that new insights can be unlocked that could create a privacy risk for the users of smart devices.

In Chapter 5, we address RQ2 by concentrating on fitness trackers from three brands, namely Fitbit, Garmin and Xiaomi. We investigate if the analysis and exploitation of the data collected by those trackers can lead to the extraction of inferences about the owners routines, health status or other sensitive information. We have used the results from the literature review we performed for this work in combination with our previous research in the thesis, and we produced a list of possible inferences that pose a threat to user privacy when using fitness trackers. Using a number of public fitness tracker datasets, as well as data that we specifically collected for this experiment from eight volunteers who wore specific fitness trackers for a period of two months, we employed statistical analysis and descriptive analytics techniques in order to analyse the data and identify patterns on the data, with the intention to identify if any particular data points or the combination of them can facilitate the elicitation of one or more of the designated inferences. The findings from this analysis were really helpful into providing insights regarding the inferences that can be extracted for each of the three fitness tracker brands under study.

As fitness trackers have completely changed how users can track and eventually monitor and evaluate their physical activity, sleep habits, and health, they have become an essential gear in users' lifestyles, since they create plenty of data that may be used for many different purposes. The range of such information accessible to fitness tracker companies is enormous, along with the possible effects from the conclusions that might be made from the fitness tracker companies using that information. The manufacturers of fitness trackers have access to a wealth of information about users' exercise habits, sleep patterns, heart rate, and other data. Through these data, they are able to identify trends and learn more about a person's general health and lifestyle habits. For example, based on variations of heart rate data, they may be able to determine if a person leads a sedentary lifestyle or performs regular physical exercise, the quantity and quality of their sleep, and also their level of stress. Furthermore, as the of majority of fitness trackers have GPS capabilities, this allows fitness tracker companies to monitor users' whereabouts during exercises or throughout the day. Advanced

Alexia Dini

sensors used in fitness trackers, such as heart rate monitors and blood oxygen level sensors, can record critical health data. Companies can exploit this information to identify potential medical concerns, like heart health issues or sleep problems. These findings raise questions about user and data privacy, as well as the appropriate handling of personal information by fitness tracker companies, even though the data collected by fitness trackers can be used in various beneficial ways.

Our experimental research results in Chapters 4 and 5 aim to provide an answer to RQ2 and at the same time can serve as a starting point for a collective effort to increase awareness among smart device owners regarding privacy risks. By guiding their attention towards protective measures, our findings aim to enhance their knowledge and shield them from potential harm. Furthermore, these insights can lead to better services for users, ultimately improving their overall experience with smart devices. Fitness tracker manufacturers must find a balance between the protection of users' privacy and data-driven inferences, so that the user trust is maintained, while the required services are provided.

10.2.3 Are the users aware of the inferences that can be made about them from their fitness trackers data?

Chapter 6 aims to address research question RQ3. For providing an answer to this research question, we conducted an online questionnaire that targeted fitness trackers users in order to gain an understanding of their concerns over their privacy when using their devices, their awareness of what data are collected by their fitness trackers and how these are being used and shared, and their awareness on the privacy risks from fitness trackers data. The chapter presents the findings of this questionnaire, through which it is highlighted that the main issue is the lack of user awareness of the inferences that can be drawn from their fitness tracker data. This survey has verified the aim of our research and supports our effort and suggestions that users must be educated about smart devices privacy risks, and also that policy makers and regulatory organisations should engage in actions aiming to increase the privacy awareness of users of smart devices in general. To that end, it is essential to provide tools and methods that enable the increase of privacy awareness.

Alexia Dini

10.2.4 Can we enhance the awareness of the users regarding the possible inferences that can be obtained from their fitness tracker data?

To address research question RQ4, Chapter 6 illustrates the methodology used as a follow up to the first questionnaire, where the same group of fitness trackers users were provided with a number of datasets from three fitness trackers brands and were asked to use them in order to interact with the PrivacyEnhAction web application. Consequently, they were asked to complete an evaluation questionnaire about the app, where they were also expected to answer similar questions to the questionnaire used in RQ3, in order to gain an understanding of whether their awareness regarding inferences has increased after using the app. In Chapters 5 and 6, our research was guided by the ambition to create a tool that will increase the users awareness in the area of fitness trackers with reference to what information can be extracted about them from the data collected and shared by their fitness trackers. Our intention was to educate the users about the possible risks and enable them to set their privacy preferences on their fitness trackers accordingly. In Chapter 9, we presented this tool, PrivacyEnhAction, and using the methodology presented in Chapter 6, we had the possibility to evaluate the impact that such a tool can have to the users' awareness.

The findings from this study assisted in understanding how users perceive data privacy, data collection and sharing, and demonstrated that there is a positive relationship between the use of a privacy awareness mechanism, like PrivacyEnhAction, and the increase of the awareness of the user about the possible privacy risks of using a fitness tracker, and as a consequence any IoT connected device. When the users are empowered with control over their privacy by making them understand the data practices of the smart devices they own, this practice adds to the strengthening of their privacy awareness. The communication of the potential privacy risks to the users and its effect to the users awareness has also been found to be important from the study, as the users privacy awareness had a positive relationship with informing them about any potential privacy risks, being in line with previous studies which give directions for the creation of privacy awareness mechanisms.

Alexia Dini

10.3 Thesis Contributions

This doctoral thesis has aimed to fill important research gaps by focusing on the privacy-aware usage of fitness trackers and smart home devices in the context of the GDPR era.

By identifying the characteristics that a user-centric GDPR-compliant privacy preserving framework in IoT should possess so that the users could be empowered with control over their personal data and privacy, by determining the inferences that can be extracted about the users from data collected from these devices, and by examining user awareness in this context, this thesis provides novel insights into an evolving and critical area of study.

In summary, the contributions of this thesis are the following:

- A list of characteristics that a user-centric GDPR-compliant privacy framework in IoT should possess in order to empower the users to be in control of their personal data and privacy. The identified characteristics can provide actionable guidance for practitioners and policymakers, while the practical implications contribute to the ongoing efforts to align IoT devices and services with privacy regulations, empower users with data control, and promote responsible data practices.
- A basis for the design and development of effective user privacy frameworks in IoT, that can be used by researchers for carrying out further research in the area, or by practitioners who can incorporate the characteristics to their platforms or systems, for providing a better protection to their users.
- A conceptual user-centric framework for IoT, “Privacy-EnhAction", that is based upon the identified list of characteristics of a user-centric GDPR-compliant privacy framework in the context of IoT. This framework provides a holistic understanding of the essential components and principles that should govern privacy frameworks in IoT environments. It represents a significant contribution to the field by offering a structured approach for researchers and practitioners to design privacy-preserving systems and services.
- A taxonomy of the inferences from smart home devices and fitness trackers. This doctoral thesis contributes experimental findings regarding the inferences that can be made from data collected from fitness trackers and smart home devices. Through the employment of machine learning techniques, descriptive analytics and a systematic literature review specific inferences have been identified that can be extracted from these smart devices. This empirical evidence adds to the existing body of knowledge by providing concrete examples and insights into the potential privacy risks and implications associated with the usage of these devices. These findings contribute to the understanding of the data inference landscape and can inform the development of mitigation strategies and privacy-enhancing measures.
- The “PrivacyEnhAction" web application. The contribution of PrivacyEnhAction in

Alexia Dini

this thesis enables users to analyse the data collected by their smart devices or fitness trackers, enhancing their awareness of privacy vulnerabilities and inferences. The tool can be utilised by individuals, privacy advocacy organisations, researchers, and device manufacturers to empower users, educate the public, advance research, and improve the privacy landscape of smart devices and fitness trackers.

- The results of a quantitative survey targeting the users of fitness trackers users for the evaluation of the level of their awareness regarding the data collected and shared by their devices and the possible inferences that can be made from their data using PrivacyEnhAction as a tool for assessing the user awareness. The identification of the lack of user awareness about the inferences that can be extracted from fitness tracker data addresses a crucial issue in the GDPR age. User awareness and education regarding privacy risks are ongoing challenges that require long-term attention. By shedding light on this issue and proposing strategies to enhance user awareness, this thesis contributes to a broader discourse on privacy education and user empowerment. The findings can inspire long-term efforts to develop educational materials, privacy literacy initiatives, and user-centric approaches that continue to empower individuals to make informed decisions about their data.
- A review of the privacy policies of a number of fitness trackers and smart home devices. By thoroughly reviewing and analysing these policies, this thesis provides insights on how data collection and sharing practices are addressed and communicated to users. This fills a significant research gap as privacy policies often serve as the primary source of information for users to understand the data practices of these devices. The results of this analysis have implications for users, privacy advocacy organisations, device manufacturers, researchers, and regulatory bodies. The findings offer valuable insights into data privacy practices and can drive improvements in privacy policies, empower users, inform advocacy efforts, and contribute to the development of privacy-conscious technologies.
- The “SpotAware" approach that classifies the text of privacy policies from the domains of fitness trackers and smart homes. This doctoral thesis makes significant contributions to the field of data privacy and user awareness, privacy advocacy, regulatory efforts, device manufacturing practices, and academic research. The “SpotAware" approach enables the extraction of GDPR user rights and inference risks from privacy policies, benefiting: (a) users who can be empowered to make informed decisions, maintain their rights and take necessary measures to protect their privacy, (b) privacy

Alexia Dini

advocacy organisations, who can assess the level of compliance with GDPR user rights and support their advocacy efforts by identifying areas of concern and advocating for stronger privacy protections and transparent practices, (c) regulatory bodies and policymakers, who can assess the compliance of privacy policies with GDPR user rights and identify potential gaps or areas that require further regulation, for the development of guidelines and policies that ensure better protection of user rights and address emerging inference risks, (d) device manufacturers, who can assess the extent to which their policies address GDPR user rights and inference risks, and be guided in improving their policies, ensuring compliance with regulations, and enhancing transparency in data collection and inference practices, and (e) researchers, who can build upon the SpotAware approach to further investigate and explore privacy policy analysis, user rights, and inference risks.

- A systematisation of inference groups that include possible inferences that could be drawn about the users from privacy policy texts. The systematisation of inference groups can serve as a valuable resource for researchers, privacy advocates, regulatory bodies, privacy-conscious users, and technology companies. It provides a comprehensive understanding of the potential inferences drawn from privacy policies, empowering stakeholders to make informed decisions, advocate for privacy rights, and drive improvements in privacy practices.
- Two annotated datasets of 133 privacy policies of smart home devices and fitness trackers. The availability of these annotated datasets benefits researchers, privacy professionals, technology companies, regulatory bodies, and consumer advocacy organisations by providing a valuable resource for analysis, policy development, compliance, and advocacy efforts.

In conclusion, this doctoral thesis makes contributions to knowledge by providing a conceptual framework, empirical insights, user awareness findings, practical implications, and methodological advancements. By identifying the characteristics of a GDPR-compliant privacy framework and proposing strategies to enhance user awareness, the thesis contributes to the development of privacy-preserving technologies, informs policy decisions, and empowers individuals to protect their privacy rights. These practical implications have the potential to positively impact the design and implementation of IoT devices, the formulation of privacy regulations, and the awareness of users regarding their data privacy. These contributions advance the understanding of privacy challenges in IoT environments, inform the development of user-centric privacy frameworks, and guide the design of privacy-enhancing technolo-

Alexia Dini

gies. By disseminating these findings, the thesis contributes to the collective knowledge of researchers, practitioners, and policymakers in the field of privacy and data protection in IoT.

10.4 Open Research Issues

The existing research in the protection of the users' privacy and personal data looks very promising, while leaving room for improvement, especially after the introduction of the GDPR. Since IoT devices can be complex and heterogeneous, the challenge of privacy is always present making the need for reliable and effective privacy protection mechanisms a top priority. Solutions are required which will enable the operation of the different domains of applications while at the same time preserving the users' privacy. Reviewing the recent literature revealed that there is not much work in the protection of the user privacy and personal data, due to the fact that the field is quite recent. There is room for more research work in the following more specific areas.

Privacy risk analysis: A third party that has already received similar data from a user can combine the personal data revelation history of the user with a privacy risks analysis of data to make inferences about that user. The absence of adequate restraints on the IoT created data and especially on how such data can be combined, makes the users lose control over their personal data. In order to cope with this challenge, solutions are needed which provide risk analysis and enhance user control over their personal data in IoT environments. This can be achieved by the integration of actions that the users can perform, such as data anonymization before sharing, a process that improves the control of the users over their personal data. Even though some studies have provided solutions towards this issue [26, 77, 274], what has not been considered is the evaluation of specific data vulnerabilities in different IoT devices. In this doctoral thesis, we have aimed to address this open research issue, by exploring the privacy risks associated with the extraction of inferences from smart home devices and fitness trackers data, proving that these inferences can reveal sensitive information about individuals, while users are not always aware about the possibility of such inferences and the extent to which their personal data are being collected, processed, and shared.

Usable user interfaces: User interfaces have become crucial in user-centric environments, like IoT. An important topic that is unfolding with IoT is the interaction of humans with their devices, through the provision of user interfaces, where the users can navigate, specify their privacy settings, communicate with the system through feedback and actions, etc. An IoT user interface should provide a personalised experience to the users, especially

Alexia Dini

since technology advances are expanding the possibilities for user interaction which generates data, and therefore a user interface becomes a perceptible part of the IoT. However, a confusing or complex interface design, with a limited visual feedback may hinder the use of available privacy options. In the reviewed papers, a very good example of user interface in terms of usability and functionality is presented in [77], where the provided interface allows the user to have a clear understanding of the associated privacy risks using colour coding for annotating them. Good examples can also be found in some other works [26, 201].

Informed consent: The process for addressing informed consent in IoT becomes a challenge, as the users are sometimes not apprehensive about the fact that their data is collected and shared. The provision of consent requests must be made in "*an intelligent and easily accessible form using clear and plain language*" [249]. The GDPR calls for the implementation of simple and informed consent, since IoT devices generate and collect huge amounts of data. The services have to comply to the GDPR requirements in order to provide quality, since if the relevant data is not provided due to uncertainty and lack of trust, this may have a negative influence on the user experience [228]. Therefore, more research effort is needed for the provision of notices to users in a clear and easy to understand design, which will enable them to provide their informed consent. A few of the works reviewed are very good examples regarding the provision of notices to users: privacy feedback is given to the user in [99], the system informs the user about possible inference risks in [288], the system provides a notice to the user regarding the risks of data sharing and also on managing those risks in [116]. Additionally, more research on the representation of data processing using a policy language, which will make the practices similar to natural language, easy for the user to understand, would be beneficial.

Context-aware user privacy preferences modelling: Another open issue is the ability to model the user's privacy preferences in IoT, where the context can change often. This has been achieved in the domain of smartphones, but not in other IoT areas. The context has to be considered very carefully to enable the user to make the correct decision. Machine learning techniques have been used in some cases, e.g. for the automatic configuration of users' privacy settings [75]. In other cases, similar techniques have been used for the recommendation of privacy settings [179], but these recommendations are static, or they are based on data sets collected from privacy-conscious and tech-savvy users, limiting the generalisation of the results [319]. Therefore, the incorporation of context-sensitive factors in the design of user-centric privacy frameworks is necessary. More research is needed to look at how machine learning and clustering techniques can be exploited in order to facilitate the specification of

Alexia Dini

preferences by learning from past activity, and also on specific techniques that can represent and monitor the elements of a given context relating to the user's privacy settings. This will enable the detection of changes in the context that call for adapting the privacy preferences of the user.

Cloud-Based IoT: Since some of the works are relevant to the integration of cloud systems and IoT, it is worth mentioning that the design of solutions for user-centric privacy in such a scenario remains an open research issue. We have not encountered such approaches in the reviewed literature. Another particularly important issue is the provision of appropriate rules and policies for ensuring that only users that have been approved can access the data in a cloud service. This is very important for the protection and preservation of users' privacy [271].

10.5 Takeaway messages

As take away messages, we urge the service providers, manufacturers and developers of smart home devices and fitness trackers to take on a number of recommendations in order to achieve a balance between the protection of the users' data privacy and the provision of the required services.

- **Service providers:** It is important that service providers implement transparency in their communication with their users, so that their data collection, sharing and processing practices are clearly communicated in the form of easy to understand privacy policies. This way, the users can make informed decisions regarding the sharing of their data. In this doctoral thesis, the important of transparency has been identified through the proposed characteristics in Chapter 3, in Characteristic CR8, "*Provide transparency*". Another important aspect that has been identified through the proposed characteristics is the need to provide data transformation, (CR2, '*Provide data transformation*), so that the users' data is protected. Appropriate techniques must be employed by the manufacturers of fitness trackers and smart home devices to minimise the risks and protect user personal data.
- **Developers:** As already mentioned in 10.4, the need for user interfaces is crucial in order for the users to specify their privacy preferences, communicate with the system through feedback and actions, and provide their informed consent. The provision of user-friendly interfaces in smart devices is crucial for empowering users be in control over their personal information and privacy. In our research, this requirement has

Alexia Dini

also been identified through a combination of characteristics, which are required in a user-friendly interface, namely CR15, “*Provide ability to users to specify their privacy preferences*, CR12, “*Provide ability to users to make informed consent choices*, and CR6, “*Provide tools for data management to users*. Additionally, it is critical to provide the means for user education and awareness in order for users to understand their data and take decisions based on their informed consent. Such methods include user-friendly interfaces, the ability for the user to provide their informed privacy preferences, and information related to their data in the form of info-graphics, tool-tips, charts, graphs, etc. This requirement has been identified through characteristic CR3, “*Provide user awareness on data collection*, as under the GDPR informed consent calls for the awareness and perception of the user of how her data are collected and shared. In the research we conducted in this doctoral thesis, we have recognised the need for the development of tools and user awareness mechanisms that can assist the users in understanding how the data created by their smart devices can be exploited for the extraction of inferences about them, and we have contributed with the PrivacyEhAction web application.

- **Developers and manufacturers:** Developers and manufacturers should work together with regulators and legislators, in order to support, maintain and further develop uniform standards for data privacy across the industry. Through the transfer of existing knowledge, experience and the sharing of best practices being used between the members of the ecosystem, a high level of data protection can be achieved and user trust in these systems and services can be increased.

We believe that by embracing these recommendations, manufacturers, service providers and developers of fitness trackers and smart home devices can assist in creating systems that will protect user data privacy, will empower the users to be in control of their data, as required by regulations such as the GDPR, and will open the door for ethical and responsible data usage in these domains.

10.6 Conclusions and Future Work

This doctoral thesis has explored the privacy risks associated with the extraction of inferences from smart home devices and fitness trackers data, as well as the level of user awareness of these risks. Through an extensive review of the literature and the development of a survey methodology, we have demonstrated that the collection, processing, and sharing of personal

Alexia Dini

data from these devices pose significant threats to individual privacy and data protection, and that users may not be fully aware of these risks.

Our research has shown that the inferences extracted from smart home devices and fitness trackers data can reveal sensitive information about individuals, such as their health status, daily routine, and lifestyle choices, among others, and that users are not always aware about the possibility of such inferences and the extent to which their personal data are being collected, processed, and shared. This lack of awareness can have serious consequences for individual privacy and data protection, and raises important privacy concerns that need to be addressed.

Overall, the findings presented in this thesis have important implications for privacy and data protection in the context of smart home devices and fitness trackers, as well as for user awareness and education. Our research highlights the need for stronger privacy regulations and standards that address the unique challenges posed by these devices, including the need for more transparent data collection and processing practices, stronger user consent mechanisms, and better technical and organisational measures to safeguard personal data. Additionally, our findings suggest the need for greater user awareness and education about the privacy risks associated with smart home devices and fitness trackers, including the use of privacy-enhancing technologies and practices.

In summary, this thesis has provided important insights into the privacy risks associated with the extraction of inferences from smart home devices and fitness trackers data, as well as user awareness of these risks. It is our hope that this work will inform and guide future research and policy initiatives aimed at protecting individual privacy and data protection, as well as improving user awareness and education in the context of emerging technologies. By addressing the issues of user privacy, awareness, and regulatory compliance, this thesis aims to empower individuals, inform policymakers, and contribute to the ongoing discourse surrounding privacy-aware usage of fitness trackers and smart home devices.

The future work plan that is based on the work presented in this thesis is as follows:

Implementation of the Privacy-EnhAction framework:. In this thesis we have focused on the second and third step of the proposed framework. Future work could focus on the implementation on the rest of the proposed steps of the framework, by developing and testing each step in a real-world IoT system. This could involve the development of tools and integrating them with smart devices to ensure that users have full control over their personal data. One tool could be the development of a negotiation mechanism between the

Alexia Dini

user and the third party where both the user and the third party will be presented with recommendations after the privacy risks analysis [26, 294] (Step 5 of Framework), or the creation a tool for the transformation of data before being released using the appropriate technique [16, 287, 288, 294, 295, 297] (Step 6 of Framework).

Further analysis of privacy policies - longitudinal and multilingual analysis:. Further work that could benefit our research performed in Chapter 8 would be the inclusion of the analysis of different regulations, which vary across different countries and regions. For example, some of the most notable ones used worldwide include (apart from the GDPR): the California Consumer Privacy Act (CCPA), the Personal Information Protection and Electronic Documents Act (PIPEDA), implemented in Canada in 2001, the Personal Data Protection Act (PDPA), implemented in Singapore in 2012, the Privacy Act, implemented in Australia in 1988 or the General Data Protection Law (LGPD), implemented in Brazil in 2020. This analyses could examine differences in policies of the same service across different jurisdictions, if there are any. Furthermore, the identification of trends in policies with the progress of time would be significantly helpful to users and regulators through a longitudinal analysis of the privacy policies. Future work in this area could also be the addition of a multilingual analysis examining the privacy policy of the same service in multiple languages in order to discover if there is any variation in terms of the metrics used [66].

Investigate the use of Large Language Models for text classification and inference identification tasks:. Large Language Models (LLMs) are deep learning models that have gained popularity in recent years for their ability to understand and generate natural language [86, 256]. Interest in LLMs is on the rise especially after the release of ChatGPT in November 2022. In this thesis, we considered multi-label classifiers for the text classification and inference identification tasks in the SpotAware methodology presented in Chapter 8. Future work could involve the investigation of the possibility of using Large Language Models for these tasks.

Alexia Dini

Bibliography

[1] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] A. Acquisti and R. Gross. Predicting social security numbers from public data. *Proceedings of the National academy of sciences*, 106(27):10975–10980, 2009.

[3] B. Ajana. Introduction: Metric culture and the over-examined life. In *Metric culture*, pages 1–9. Emerald Publishing Limited, 2018.

[4] H. Aksu, L. Babun, M. Conti, G. Tolomei, and A. S. Uluagac. Advertising in the iot era: Vision and challenges. *IEEE Communications Magazine*, 56(11):138–144, 2018.

[5] A. Aktypi, J. R. Nurse, and M. Goldsmith. Unwinding ariadne's identity thread: Privacy risks with fitness trackers and online social networks. In *Proceedings of the 2017 on Multimedia Privacy and Security*, pages 1–11. 2017.

[6] S. S. Al-Fedaghi. The right to be let alone and private information. In *Enterprise Information Systems VII*, pages 157–166. Springer, 2006.

[7] Z. Al-Makhadmeh and A. Tolba. Utilizing iot wearable medical device for heart disease prediction using higher order boltzmann model: A classification approach. *Measurement*, 147:106815, 2019.

[8] K. Alanezi and S. Mishra. A privacy negotiation mechanism for the internet of things. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/-DataCom/CyberSciTech)*, pages 512–519. IEEE, 2018.

[9] Alcatel. Privacy Policy. `https://www.alcatelmobile.com/eu/privacy/`, 2011. [Online; Accessed January 2023].

[10] L. Alhalabi, M. J. Singleton, A. O. Oseni, A. J. Shah, Z.-M. Zhang, and E. Z. Soliman. Relation of higher resting heart rate to risk of cardiovascular versus noncardiovascular death. *The American journal of cardiology*, 119(7):1003–1007, 2017.

[11] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf. A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, pages 3–21, 2020.

[12] A. Alqhatani and H. R. Lipford. Exploring the design space of sharing and privacy mechanisms in wearable fitness platforms. In *Workshop on Usable Security and Privacy (USEC)*, volume 7, 2021.

[13] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie. {PolicyLint}: Investigating internal privacy policy contradictions on google play. In *28th USENIX security symposium (USENIX security 19)*, pages 585–602, 2019.

[14] S. Arca and R. Hewett. Privacy protection in smart health. In *Proceedings of the 11th International Conference on Advances in Information Technology*, pages 1–8, 2020.

[15] E. Arfelt, D. Basin, and S. Debois. Monitoring the gdpr. In *European Symposium on Research in Computer Security*, pages 681–699. Springer, 2019.

[16] F. R. Asl, F. Chiang, W. He, and R. Samavi. Privacy aware web services in the cloud. In *2017 IEEE Conference on Communications and Network Security (CNS)*, pages 458–466. IEEE, 2017.

[17] Asus. Terms of Use Notice / Privacy Policy. `https://www.asus.com/Terms_of_Use_Notice_Privacy_Policy/Privacy_Policy/`, 2022. [Online; Accessed April 2023].

[18] M. Bada and B. von Solms. A cybersecurity guide for using fitness devices. *arXiv preprint arXiv:2105.02933*, 2021.

[19] P. Bahirat, Y. He, A. Menon, and B. Knijnenburg. A data-driven approach to developing iot privacy-setting interfaces. In *23rd International Conference on Intelligent User Interfaces*, pages 165–176. ACM, 2018.

[20] P. Bai, S. Kumar, K. Kumar, O. Kaiwartya, M. Mahmud, and J. Lloret. Gdpr compliant data storage and sharing in smart healthcare system: a blockchain-based solution. *Electronics*, 11(20):3311, 2022.

[21] M. Baig. Can a Fitness Tracker Detect Diabetes? `https://precisiondrivenhealth.com/can-a-fitness-tracker-detect-diabetes/`, 2017. [Online; Accessed April 2023].

[22] F. Bakalov, M.-J. Meurs, B. König-Ries, B. Sateli, R. Witte, G. Butler, and A. Tsang. An approach to controlling user models and personalization effects in recommender systems. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 49–56, 2013.

[23] S. Balakrishnan, H. Vasudavan, and R. K. Murugesan. Smart home technologies: A preliminary review. In *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, pages 120–127, 2018.

[24] V. E. Balas, V. K. Solanki, R. Kumar, and M. A. R. Ahad. *A Handbook of Internet of Things in Biomedical and Cyber Physical System.* Springer, 2020.

[25] N. Balta-Ozkan, R. Davidson, M. Bicket, and L. Whitmarsh. Social barriers to the adoption of smart homes. *Energy Policy*, 63:363–374, 2013.

[26] M. Barhamgi, C. Perera, C. Ghedira, and D. Benslimane. User-centric privacy engineering for the internet of things. *IEEE Cloud Computing*, 5(5):47–57, 2018.

[27] M. Barhamgi, M. Yang, C.-M. Yu, Y. Yu, A. K. Bandara, D. Benslimane, and B. Nuseibeh. Enabling end-users to protect their privacy. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 905–907, 2017.

[28] B. R. Barricelli, E. Casiraghi, J. Gliozzo, A. Petrini, and S. Valtolina. Human digital twin for fitness management. *Ieee Access*, 8:26637–26664, 2020.

[29] S. Becher, A. Gerl, and B. Meier. Don t forget the user: From user preferences to personal privacy policies. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 774–778. IEEE, 2020.

Alexia Dini

[30] V. Beltran, A. F. Skarmeta, and P. M. Ruiz. An arm-compliant architecture for user privacy in smart cities: Smartiequality by design in the iot. *Wireless Communications and Mobile Computing*, 2017, 2017.

[31] F. Z. Berrehili and A. Belmekki. Privacy preservation in the internet of things. In *International Symposium on Ubiquitous Networking*, pages 163–175. Springer, 2016.

[32] J. Bhatia, M. C. Evans, and T. D. Breaux. Identifying incompleteness in privacy policy goals using semantic frames. *Requirements Engineering*, 24(3):291–313, 2019.

[33] Z. Bilgin, E. Tomur, M. A. Ersoy, and E. U. Soykan. Statistical appliance inference in the smart grid by machine learning. In *2019 IEEE 30th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops)*, pages 1–7. IEEE, 2019.

[34] J. Blasco, T. M. Chen, H. K. Patil, and D. Wolff. Wearables security and privacy. In *Mission-Oriented Sensor Networks and Systems: Art and Science*, pages 351–380. Springer, 2019.

[35] P. W. Blog. The easiest way to interpret clustering result.

[36] F. Blow, Y.-H. Hu, and M. Hoppa. A study on vulnerabilities and threats to wearable devices. In *Journal of The Colloquium for Information Systems Security Education*, volume 7, pages 7–7, 2020.

[37] A. Bonaquist, M. Grehan, O. Haines, J. Keogh, T. Mullick, N. Singh, S. Shaaban, A. Radovic, and A. Doryab. An automated machine learning pipeline for monitoring and forecasting mobile health data. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE, 2021.

[38] F. W. Booth, C. K. Roberts, and M. J. Laye. Lack of exercise is a major cause of chronic diseases. *Comprehensive physiology*, 2(2):1143, 2012.

[39] S. T. M. Bourobou and Y. Yoo. User activity recognition in smart homes using pattern clustering applied to temporal ann algorithm. *Sensors*, 15(5):11953–11971, 2015.

[40] M. Bourreau. Google - Fitbit. `https://voxeu.org/article/googlefitbit-will-monetise-health-data-and-harm-consumers`, 2020. [Online; Accessed April 2023].

[41] L. Brandeis and S. Warren. The right to privacy. *Harvard law review*, 4(5):193–220, 1890.

[42] G. Broenink, J.-H. Hoepman, C. v. Hof, R. Van Kranenburg, D. Smits, and T. Wisman. The privacy coach: Supporting customer privacy in the internet of things. *arXiv preprint arXiv:1001.4459*, 2010.

[43] J. Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2):29–40, 2013.

[44] L. Bufalieri, M. La Morgia, A. Mei, and J. Stefa. Gdpr: when the right to access personal data becomes a threat. In *2020 IEEE International Conference on Web Services (ICWS)*, pages 75–83. IEEE, 2020.

[45] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110, 2021.

Alexia Dini

[46] B. Carminati, P. Colombo, E. Ferrari, and G. Sagirlar. Enhancing user control on personal data usage in internet of things ecosystems. In *2016 IEEE International Conference on Services Computing (SCC)*, pages 291–298. IEEE, 2016.

[47] A. Cavoukian and J. Jonas. *Privacy by design in the age of big data*. Information and Privacy Commissioner of Ontario, Canada, 2012.

[48] CEOToday. Is Data The New Gold? `https://www.ceotodaymagazine.com/2018/04/is-data-the-new-gold/`, 2020. [Online; Accessed August 2022].

[49] N. Challa, S. Yu, and S. Kunchakarra. Wary about wearables: Potential for the exploitation of wearable health technology through employee discrimination and sales to third parties. *Intersect: The Stanford Journal of Science, Technology, and Society*, 10(3), 2017.

[50] R. Chamarajnagar and A. Ashok. Privacy invasion through smarthome iot sensing. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2019.

[51] C. Chang, H. Li, Y. Zhang, S. Du, H. Cao, and H. Zhu. Automated and personalized privacy policy extraction under gdpr consideration. In *International Conference on Wireless Algorithms, Systems, and Applications*, pages 43–54. Springer, 2019.

[52] L. Chang, J. Lu, J. Wang, X. Chen, D. Fang, Z. Tang, P. Nurmi, and Z. Wang. Sleepguard: Capturing rich sleep information using smartwatch sensing data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–34, 2018.

[53] P. Chaudhari and S. Sane. Multilabel classification exploiting coupled label similarity with feature selection. *IJCA ETC*, 142, 2016.

[54] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy. Non-intrusive occupancy monitoring using smart meters. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8, 2013.

[55] L. F. Chen and R. Ismail. Information technology program students' awareness and perceptions towards personal data protection and privacy. In *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pages 434–438. IEEE, 2013.

[56] R. Chen and Y. Tong. A two-stage method for solving multi-resident activity recognition in smart environments. *Entropy*, 16(4):2184–2203, 2014.

[57] S. Chen. Interpretation of multi-label classification models using shapley values. *arXiv preprint arXiv:2104.10505*, 2021.

[58] Y. Chen and C. Shen. Performance analysis of smartphone-sensor behavior for human activity recognition. *Ieee Access*, 5:3095–3110, 2017.

[59] E. A. Cherman, N. Spolaôr, J. Valverde-Rebaza, and M. C. Monard. Lazy multi-label learning algorithms based on mutuality strategies. *Journal of Intelligent & Robotic Systems*, 80:261–276, 2015.

Alexia Dini

[60] C. Chhetri and V. G. Motti. Eliciting privacy concerns for smart home devices from a user centered perspective. In *Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings 14*, pages 91–101. Springer, 2019.

[61] S. Chitkara, N. Gothoskar, S. Harish, J. I. Hong, and Y. Agarwal. Does this app really need my location? context-aware privacy management for smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–22, 2017.

[62] J. Y. Cho, D. Ko, and B. G. Lee. Strategic approach to privacy calculus of wearable device user regarding information disclosure and continuance intention. *KSII Transactions on Internet and Information Systems (TIIS)*, 12(7):3356–3374, 2018.

[63] Collins. Definition of inference. `https://www.collinsdictionary.com/dictionary/english/inference`, 2023. [Online; Accessed April 2023].

[64] I. Consulting. GDPR Personal Data. `https://gdpr-info.eu/issues/personal-data/`, 2018. [Online; Accessed April 2023].

[65] I. Consulting. GDPR Rights. `https://gdpr-info.eu/chapter-3/`, 2018. [Online; Accessed August 2022].

[66] G. Contissa, K. Drazewski, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Pałka, G. Sartor, and P. Torroni. Gdpr privacy policies in claudette: Challenges of omission, context and multilingualism. 2019.

[67] J. Cook. Inferring religion. `https://dzone.com/articles/inferring-personal-information-from-fitness-data`, 2021. [Online; Accessed April 2023].

[68] M. T. Cooney, E. Vartiainen, T. Laakitainen, A. Juolevi, A. Dudina, and I. M. Graham. Elevated resting heart rate is an independent risk factor for cardiovascular disease in healthy men and women. *American heart journal*, 159(4):612–619, 2010.

[69] R. Costa and A. Pinto. A framework for the secure storage of data generated in the iot. In *Ambient Intelligence-Software and Applications*, pages 175–182. Springer, 2015.

[70] E. Costante, J. den Hartog, and M. Petković. What websites know about you. In *International Workshop on Data Privacy Management, International Workshop on Autonomous and Spontaneous Security*, pages 146–159. Springer, 2013.

[71] M. Cremonini, C. Braghin, and C. A. Ardagna. Privacy on the internet. In *Computer and information security handbook*, pages 739–753. Elsevier, 2013.

[72] Cubot. Privacy Policy. `https://shop.cubot.net/pages/privacy-policy`, 2020. [Online; Accessed January 2023].

[73] B. Custers and A.-S. Heijne. The right of access in automated decision-making: The scope of article 15 (1)(h) gdpr in theory and practice. *Computer Law & Security Review*, 46:105727, 2022.

[74] A. Das, M. Degeling, D. Smullen, and N. Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing*, 17(3):35–46, 2018.

Alexia Dini

[75] A. Das, M. Degeling, X. Wang, J. Wang, N. Sadeh, and M. Satyanarayanan. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1387–1396. IEEE, 2017.

[76] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra. Uncovering privacy leakage in ble network traffic of wearable fitness trackers. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications*, pages 99–104, 2016.

[77] S. J. De and D. Le Métayer. Privacy risk analysis to enable informed privacy settings. In *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 95–102. IEEE, 2018.

[78] Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. S. Pentland. Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 48–55. Springer, 2013.

[79] M. Debnath, P. K. Tripathi, and R. Elmasri. K-dbscan: Identifying spatial clusters with differing density levels. In *2015 International workshop on data mining with industrial applications (DMIA)*, pages 51–60. IEEE, 2015.

[80] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz. We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*, 2018.

[81] M. F. Dennedy, J. Fox, and T. R. Finneran. *The privacy engineer's manifesto: getting from policy to code to QA to value.* Springer Nature, 2014.

[82] R. Ding, H. Zhong, J. Ma, X. Liu, and J. Ning. Lightweight privacy-preserving identity-based verifiable iot-based health storage system. *IEEE Internet of Things Journal*, 6(5):8393–8405, 2019.

[83] A. Dini Kounoudes, G. M. Kapitsaki, and I. Katakis. Enhancing user awareness on inferences obtained from fitness trackers data. *User Modeling and User-Adapted Interaction*, pages 1–48, 2023.

[84] A. Dini Kounoudes, G. M. Kapitsaki, and M. Milis. Towards considering user privacy preferences in smart water management. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 209–212, 2019.

[85] W. Dong, B. Lepri, and A. Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143, 2011.

[86] K. Drazewski, A. Galassi, A. JABŁONOWSKA, F. Lagioia, M. Lippi, H.-W. MICKLITZ, G. Sartor, G. Tagiuri, and P. Torroni. A corpus for multilingual analysis of online terms of service. Association for Computational Linguistics, 2021.

[87] R. Y. Du, O. Netzer, D. A. Schweidel, and D. Mitra. Capturing marketing information to fuel growth. *Journal of Marketing*, 85(1):163–183, 2021.

[88] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

Alexia Dini

[89] Ecobee. Privacy Policy and Terms of Use. https://www.ecobee.com/en-us/privacy-policy/, 2022. [Online; Accessed April 2023].

[90] G. Eibl and D. Engel. Influence of data granularity on smart meter privacy. *IEEE Transactions on Smart Grid*, 6(2):930–939, 2014.

[91] F. Els and L. Cilliers. A privacy management framework for personal electronic health records. *African Journal of Science, Technology, Innovation and Development*, 10(6):725–734, 2018.

[92] EU. EU data protection rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en, 2018. [Online; Accessed March 2023].

[93] EU. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02016R0679-20160504&from=EN, 2018. [Online; Accessed March 2023].

[94] B. Faber, G. C. Michelet, N. Weidmann, R. R. Mukkamala, and R. Vatrapu. Bpdims: A blockchain-based personal data and identity management system. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[95] M. Fahim and A. Sillitti. Analyzing load profiles of energy consumption to infer household characteristics using smart meters. *Energies*, 12(5):773, 2019.

[96] J. Fan, Q. Li, and G. Cao. Privacy disclosure through smart meters: Reactive power based attack and defense. In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 13–24. IEEE, 2017.

[97] Y. Fang, Y. Liu, C. Huang, and L. Liu. Fastembed: Predicting vulnerability exploitation possibility based on ensemble machine learning algorithm. *Plos one*, 15(2):e0228439, 2020.

[98] M. Fernström, U. Fernberg, G. Eliason, and A. Hurtig-Wennlöf. Aerobic fitness is associated with low cardiovascular disease risk: the impact of lifestyle on early risk factors for atherosclerosis in young healthy swedish individuals–the lifestyle, biomarker, and atherosclerosis study. *Vascular health and risk management*, 13:91, 2017.

[99] H. S. Fhom, N. Kuntze, C. Rudolph, M. Cupelli, J. Liu, and A. Monti. A user-centric privacy manager for future energy systems. In *2010 International Conference on Power System Technology*, pages 1–7. IEEE, 2010.

[100] K. Fietkiewicz and A. Ilhan. Fitness tracking technologies: Data privacy doesnt matter? the (un) concerns of users, former users, and non-users. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[101] I. Figueirêdo, L. L. N. Guarieiro, and E. G. S. Nascimento. Multivariate real time series data using six unsupervised machine learning algorithms. In *Anomaly Detection-Recent Advances, Issues and Challenges*. IntechOpen, 2020.

[102] Forbes. Data Is The New Gold. https://www.forbesafrica.com/technology/2019/07/18/data-is-the-new-gold/, 2019. [Online; Accessed February 2023].

Alexia Dini

[103] W. E. Forum. Data is the new gold. This is how it can benefit everyone while harming no one. `https://bit.ly/3eazKmm`, 2020. [Online; Accessed April 2023].

[104] N. Foukia, D. Billard, and E. Solana. Pisces: A framework for privacy by design in iot. In *Privacy, Security and Trust (PST), 2016 14th Annual Conference on*, pages 706–713. IEEE, 2016.

[105] N. Fourberg, T. Serpil, L. Wiewiorra, I. GODLOVITCH, A. DE STREEL, H. Jacquemin, H. Jordan, N. Madalina, F. JACQUES, M. LEDGER, et al. Online advertising: the impact of targeted advertising on advertisers, market access and consumer choice. 2021.

[106] C. Fried. The value of life. *Harv. L. Rev.*, 82:1415, 1968.

[107] R. Furberg, J. Brinton, M. Keating, and A. Ortiz. Crowd-sourced fitbit datasets 03.12. 2016-05.12. 2016, 2016. [Online; Accessed April 2023].

[108] Furbo. Privacy Policy. `https://furbo.com/uk/pages/privacy-policy?gclid=EAIaIQobChMI8peFg_rV_AIVRqjVCh3HWgFsEAAYASAAEgKnW_D_BwE`, 2020. [Online; Accessed January 2023].

[109] S. Gabriele and S. Chiasson. Understanding fitness tracker users' security and privacy knowledge, attitudes and behaviours. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[110] GDPR. GDPR personal data what information does this cover? `https://www.gdpreu.org/the-regulation/key-concepts/personal-data/`, 2023. [Online; Accessed March 2023].

[111] J. Geer and R. Gellert. *The GDPR: An Empowering and Protective Mechanism for Childrens Internet Connected Toy Use*. PhD thesis, Masters thesis). Tilburg University, Department of Law and Technology. Tilburg, 2018.

[112] T. Gerety. Redefining privacy. *Harv. CR-CLL Rev.*, 12:233, 1977.

[113] A. Gerl and B. Meier. The layered privacy language art. 12–14 gdpr extension–privacy enhancing user interfaces. *Datenschutz und Datensicherheit-DuD*, 43(12):747–752, 2019.

[114] M. Goddard. The eu general data protection regulation (gdpr): European regulation that has a global impact. *International Journal of Market Research*, 59(6):703–705, 2017.

[115] A. Goldsteen, S. Garion, S. Nadler, N. Razinkov, Y. Moatti, and P. Ta-Shma. Brief announcement: A consent management solution for enterprises. In *International Conference on Cyber Security Cryptography and Machine Learning*, pages 189–192. Springer, 2017.

[116] P. Grace and M. Surridge. Towards a model of user-centered privacy preservation. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*, page 91. ACM, 2017.

[117] C. Gross, W. Wenner, and R. Lackes. Using wearable fitness trackers to detect covid-19?! In *International Conference on Business Informatics Research*, pages 51–65. Springer, 2021.

Alexia Dini

[118] J. Gudgel. Objects of concern? risks, rewards and regulation in the'internet of things'. *Risks, Rewards and Regulation in the'Internet of Things'(April 29, 2014)*, 2014.

[119] Gunicorn. Gunicorn. `https://gunicorn.org/`, 2023. [Online; Accessed April 2023].

[120] N. Guntamukkala, R. Dara, and G. Grewal. A machine-learning based approach for measuring the completeness of online privacy policies. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 289–294. IEEE, 2015.

[121] GVR. Wearable Technology Market Share Trends Report, 2030. `https://www.grandviewresearch.com/industry-analysis/wearable-technology-market`, 2023. [Online; Accessed April 2023].

[122] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145, 2001.

[123] R. E. Hamdani, M. Mustapha, D. R. Amariles, A. Troussel, S. Meeùs, and K. Krasnashchok. A combined rule-based and machine learning approach for automated gdpr compliance checking. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 40–49, 2021.

[124] F. Hantke and A. Dewald. How can data from fitness trackers be obtained and analyzed with a forensic approach? In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 500–508. IEEE, 2020.

[125] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, 2018.

[126] J. He, Q. Xiao, P. He, and M. S. Pathan. An adaptive privacy protection method for smart home environments using supervised learning. *Future Internet*, 9(1):7, 2017.

[127] Healthline. Healthline. `https://www.healthline.com/`, 2023. [Online; Accessed April 2023].

[128] A. Henriksen, M. H. Mikalsen, A. Z. Woldaregay, M. Muzny, G. Hartvigsen, L. A. Hopstock, S. Grimsgaard, et al. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. *Journal of medical Internet research*, 20(3):e9157, 2018.

[129] M. Henze, L. Hermerschmidt, D. Kerpen, R. Häußling, B. Rumpe, and K. Wehrle. User-driven privacy enforcement for cloud-based services in the internet of things. In *2014 International Conference on Future Internet of Things and Cloud*, pages 191–196. IEEE, 2014.

[130] M. Henze, L. Hermerschmidt, D. Kerpen, R. Häußling, B. Rumpe, and K. Wehrle. A comprehensive approach to privacy in the cloud-based internet of things. *Future Generation Computer Systems*, 56:701–718, 2016.

[131] J. L. Hicks, T. Althoff, R. Sosic, P. Kuhar, B. Bostjancic, A. C. King, J. Leskovec, and S. L. Delp. Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ digital medicine*, 2(1):1–12, 2019.

Alexia Dini

[132] Hive. Privacy Policy. https://www.hivehome.com/privacy, 2022. [Online; Accessed January 2023].

[133] G. Högström, A. Nordström, and P. Nordström. Aerobic fitness in late adolescence and the risk of early death: a prospective cohort study of 1.3 million swedish men. *International journal of epidemiology*, 45(4):1159–1168, 2016.

[134] E. Horvitz and D. Mulligan. Data, privacy, and the greater good. *Science*, 349(6245):253–255, 2015.

[135] H. Hosseini, M. Degeling, C. Utz, and T. Hupperich. Unifying privacy policy detection. *Proc. Priv. Enhancing Technol.*, 2021(4):480–499, 2021.

[136] Y. Huang and L. Li. Naive bayes classification algorithm based on small sample set. In *2011 IEEE International conference on cloud computing and intelligence systems*, pages 34–39. IEEE, 2011.

[137] S. Hunter and S. C. Robson. Adaptation of the maternal heart in pregnancy. *British heart journal*, 68(6):540, 1992.

[138] A. Ilhan and K. J. Fietkiewicz. Data privacy-related behavior and concerns of activity tracking technology users from germany and the usa. *Aslib Journal of Information Management*, 2020.

[139] I. Ioannidou and N. Sklavos. On general data protection regulation vulnerabilities and privacy issues, for wearable devices and fitness tracking applications. *Cryptography*, 5(4):29, 2021.

[140] U. Iqbal, P. N. Bahrami, R. Trimananda, H. Cui, A. Gamero-Garrido, D. Dubois, D. Choffnes, A. Markopoulou, F. Roesner, and Z. Shafiq. Your echos are heard: Tracking, profiling, and ad targeting in the amazon smart speaker ecosystem. *arXiv preprint arXiv:2204.10920*, 2022.

[141] J. Jones and J. Seladi-Schulman. Causes of slow heart rate. https://www.healthline.com/health/slow-heart-rate#causes, 2021. [Online; Accessed April 2023].

[142] J. Joy, M. Le, and M. Gerla. Locationsafe: Granular location privacy for iot devices. In *Proceedings of the Eighth Wireless of the Students, by the Students, and for the Students Workshop*, pages 39–41, 2016.

[143] G. Jung, H. Lee, A. Kim, and U. Lee. Too much information: assessing privacy risks of contact trace data disclosure on people with covid-19 in south korea. *Frontiers in public health*, 8:305, 2020.

[144] D. W. Kaiser, R. A. Harrington, and M. P. Turakhia. Wearable fitness trackers and heart disease. *JAMA cardiology*, 1(2):239–239, 2016.

[145] H. Kang and E. H. Jung. The smart wearables-privacy paradox: A cluster analysis of smartwatch users. *Behaviour & Information Technology*, 40(16):1755–1768, 2021.

[146] Y.-S. Kao, K. Nawata, and C.-Y. Huang. An exploration and confirmation of the factors influencing adoption of iot-based wearable fitness trackers. *International journal of environmental research and public health*, 16(18):3227, 2019.

Alexia Dini

[147] A. Karami and R. Johansson. Choosing dbscan parameters automatically using differential evolution. *International Journal of Computer Applications*, 91(7):1–11, 2014.

[148] A. Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda, 2017.

[149] G. Kaur and E. N. Oberai. A review article on naive bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing*, 3(10):864–868, 2014.

[150] A. Kazlouski, T. Marchioro, H. Manifavas, and E. Markatos. Do you know who is talking to your wearable smartband? *Integrated Citizen Centered Digital Health and Social Care: Citizens as Data Producers and Service co-Creators*, 275:142, 2020.

[151] M. Keshavarz and M. Anwar. Towards improving privacy control for smart homes: A privacy decision framework. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–3. IEEE, 2018.

[152] J. W. Kim, S.-M. Moon, S.-u. Kang, and B. Jang. Effective privacy-preserving collection of health data from a users wearable device. *Applied Sciences*, 10(18):6396, 2020.

[153] W. Kleiminger, C. Beckel, and S. Santini. Household occupancy monitoring using electricity meters. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 975–986, 2015.

[154] W. Kleiminger, C. Beckel, T. Staake, and S. Santini. Occupancy detection from electricity consumption data. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8, 2013.

[155] P. Koehn, J. Schroeder, and L. Valiukas. sentence-splitter 1.4. `https://pypi.org/project/sentence-splitter/`, 2023. [Online; Accessed April 2023].

[156] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[157] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[158] A. D. Kounoudes. PrivacyEnhaction Evaluation Questionnaire. `https://forms.gle/KCJ2xx23quK4A8wk8`, 2022. Online; Accessed March 2023].

[159] A. D. Kounoudes. Questionnaire on fitness trackers user privacy concerns. `https://forms.gle/uzVzVhew2Jq3XeAS9`, 2022. [Online; Accessed March 2023].

[160] A. D. Kounoudes and G. M. Kapitsaki. A mapping of iot user-centric privacy preserving approaches to the gdpr. *Internet of Things*, 11:100179, 2020.

[161] A. D. Kounoudes, G. M. Kapitsaki, I. Katakis, and M. Milis. User-centred privacy inference detection for smart home devices. In *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*, pages 210–218. IEEE, 2021.

Alexia Dini

[162] J. Kröger. Unexpected inferences from sensor data: a hidden privacy threat in the internet of things. In *IFIP International Internet of Things Conference*, pages 147–159. Springer, 2018.

[163] J. L. Kröger, O. H.-M. Lutz, and F. Müller. What does your gaze reveal about you? on the privacy implications of eye tracking. In *IFIP International Summer School on Privacy and Identity Management*, pages 226–241. Springer, 2019.

[164] J. L. Kröger, O. H.-M. Lutz, and P. Raschke. Privacy implications of voice and speech analysis–information disclosure by inference. In *IFIP International Summer School on Privacy and Identity Management*, pages 242–258. Springer, 2019.

[165] J. L. Kröger, O. H.-M. Lutz, and P. Raschke. Privacy implications of voice and speech analysis–information disclosure by inference. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pages 242–258, 2020.

[166] J. L. Kröger, P. Raschke, and T. R. Bhuiyan. Privacy implications of accelerometer data: a review of possible inferences. In *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, pages 81–87, 2019.

[167] J. L. Kröger, P. Raschke, J. P. Campbell, and S. Ullrich. Surveilling the gamers: Privacy impacts of the video game industry. *Available at SSRN 3881279*, 2021.

[168] B. Krzanich. Data is the New Oil in the Future of Automated Driving. `https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving/#gs.4h16m3`, 2016. [Online; Accessed February 2023].

[169] S. Kumar, N. Kumar, A. Dev, and S. Naorem. Movie genre classification using binary relevance, label powerset, and machine learning classifiers. *Multimedia Tools and Applications*, 82(1):945–968, 2023.

[170] M. R. Langley. Hide your health: addressing the new privacy problem of consumer wearables. *Geo. LJ*, 103:1641, 2014.

[171] D. Le Métayer, V. Morel, and M. Cunche. *A Generic Information and Consent Framework for the IoT*. PhD thesis, Inria, 2018.

[172] L. Lee, J. Lee, S. Egelman, and D. Wagner. Information disclosure concerns in the age of wearable computing. In *NDSS Workshop on Usable Security (USEC)*, volume 1, pages 1–10, 2016.

[173] M. Lehto and M. Lehto. Health information privacy of activity trackers. In *European Conference on Cyber Warfare and Security*, pages 243–251. Academic Conferences International Limited, 2017.

[174] J. Li and S. Brewer. A performance comparison of unsupervised machine learning algorithms for clustering water depth datasets at urban drainage systems. 2020.

[175] T. Li, C. Zhang, and S. Zhu. Empirical studies on multi-label classification. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, pages 86–92. IEEE, 2006.

Alexia Dini

[176] D. Liciotti, M. Bernardini, L. Romeo, and E. Frontoni. A sequential deep learning application for recognising human activities in smart homes. *Neurocomputing*, 396:501–513, 2020.

[177] R. Liepin, G. Contissa, K. Drazewski, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Palka, G. Sartor, P. Torroni, et al. Gdpr privacy policies in claudette: Challenges of omission, context and multilingualism. In *CEUR WORKSHOP PROCEEDINGS*, volume 2385. CEUR-WS, 2019.

[178] H. R. Lipford, M. Tabassum, P. Bahirat, Y. Yao, and B. P. Knijnenburg. Privacy and the internet of things. In *Modern Socio-Technical Perspectives on Privacy*, pages 233–264. Springer, Cham, 2022.

[179] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 27–41, 2016.

[180] D. Liu, C. Wu, L. Yang, X. Zhao, and Q. Sun. The development of privacy protection standards for smart home. *Wireless Communications and Mobile Computing*, 2022, 2022.

[181] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang. Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13. In *Proceedings of the Web Conference 2021*, pages 2154–2164, 2021.

[182] X. Liu, M. Chen, C. Tan, X. Zhang, and W. Yang. Automatic stem mapping using single-scan terrestrial laser scanning data and mean shift clustering. In *IOP Conference Series: Earth and Environmental Science*, volume 865, page 012015. IOP Publishing, 2021.

[183] B. Lovejoy. Smartphone and smartwatch data led husband to confess to murdering his wife. `https://9to5mac.com/2021/06/18/smartphone-and-smartwatch-data-murder/`, 2021. [Online; Accessed March 2023].

[184] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, and A. Bahamonde. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1:303–313, 2012.

[185] A. Lukács. What is privacy? the history and definition of privacy. 2016.

[186] K. Maganti, V. H. Rigolin, M. E. Sarano, and R. O. Bonow. Valvular heart disease: diagnosis and management. In *Mayo Clinic Proceedings*, volume 85, pages 483–500. Elsevier, 2010.

[187] I. Makhdoom, I. Zhou, M. Abolhasan, J. Lipman, and W. Ni. Privysharing: A blockchain-based framework for privacy-preserving and secure data sharing in smart cities. *Computers & Security*, 88:101653, 2020.

[188] J. M. Mangrum and J. P. DiMarco. The evaluation and management of bradycardia. *New England Journal of Medicine*, 342(10):703–709, 2000.

[189] D. Marikyan, S. Papagiannidis, and E. Alamanos. A systematic review of the smart home literature: A user perspective. *Technological Forecasting and Social Change*, 138:139–154, 2019.

Alexia Dini

[190] M.-O. Mario. Human activity recognition based on single sensor square hv acceleration images and convolutional neural networks. *IEEE Sensors Journal*, 19(4):1487–1498, 2018.

[191] K. Masuch, M. Greve, and S. Trang. Fitness first or safety first? examining adverse consequences of privacy seals in the event of a data breach. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 3871, 2021.

[192] A. Mayle, N. H. Bidoki, S. Masnadi, L. Boeloeni, and D. Turgut. Investigating the value of privacy within the internet of things. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pages 1–6. IEEE, 2017.

[193] J. McCarthy. Arthur Samuel: Pioneer in Machine Learning. `http://infolab.stanford.edu/pub/voy/museum/samuel.html`, 2019. [Online; Accessed April 2023].

[194] B. McDonald, P. Pudney, and J. Rong. Pattern recognition and segmentation of smart meter data. *ANZIAM Journal*, 54:M105–M150, 2012.

[195] E. McGowan. Here's what your Fitbit knows about you. `https://blog.avast.com/what-fitbit-knows-about-you-avast`, 2021. [Online; Accessed February 2023].

[196] Ü. Meteriz, N. F. Yıldıran, and A. Mohaisen. You can run, but you cannot hide: Using elevation profiles to breach location privacy through trajectory prediction. *arXiv preprint arXiv:1910.09041*, 2019.

[197] M. Mirmomeni, T. Fazio, S. von Cavallar, and S. Harrer. From wearables to thinkables: artificial intelligence-enabled sensors for health monitoring. In *Wearable Sensors*, pages 339–356. Elsevier, 2021.

[198] B. Mittelstadt. Designing the health-related internet of things: ethical principles and guidelines. *Information*, 8(3):77, 2017.

[199] M. Mohzary, S. Tadisetty, and K. Ghazinour. A privacy protection layer for wearable devices. In *Foundations and Practice of Security: 12th International Symposium, FPS 2019, Toulouse, France, November 5–7, 2019, Revised Selected Papers*, volume 12056, page 363. Springer Nature, 2020.

[200] R. Molich and J. Nielsen. Improving a human-computer dialogue. *Communications of the ACM*, 33(3):338–348, 1990.

[201] V. Morel, M. Cunche, and D. Le Métayer. A generic information and consent framework for the iot. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 366–373. IEEE, 2019.

[202] N. Mousavi Nejad, S. Scerri, and J. Lehmann. Knight: Mapping privacy policies to gdpr. In *European Knowledge Acquisition Workshop*, pages 258–272. Springer, 2018.

[203] MyKronoz. Privacy Policy. `https://www.mykronoz.com/eu/en/privacy-policy/`, 2018. [Online; Accessed January 2023].

Alexia Dini

[204] M. Nagai, S. Hoshide, and K. Kario. Sleep duration as a risk factor for cardiovascular disease-a review of the recent literature. *Current cardiology reviews*, 6(1):54–61, 2010.

[205] U. Nations. Universal Declaration of Human Rights. `https://www.un.org/sites/un2.un.org/files/2021/03/udhr.pdf`, 2021. [Online; Accessed April 2023].

[206] R. Neisse, G. Baldini, G. Steri, and V. Mahieu. Informed consent in internet of things: The case study of cooperative intelligent transport systems. In *2016 23rd International Conference on Telecommunications (ICT)*, pages 1–5. IEEE, 2016.

[207] R. Neisse, G. Baldini, G. Steri, Y. Miyake, S. Kiyomoto, and A. R. Biswas. An agent-based framework for informed consent in the internet of things. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, pages 789–794. IEEE, 2015.

[208] R. Neisse, G. Steri, I. N. Fovino, and G. Baldini. Seckit: a model-based security toolkit for the internet of things. *computers & security*, 54:60–76, 2015.

[209] L. Nemec Zlatolas, N. Feher, and M. Hölbl. Security perception of iot devices in smart homes. *Journal of Cybersecurity and Privacy*, 2(1):65–73, 2022.

[210] J. R. Nurse, A. Atamli, and A. Martin. Towards a usable framework for modelling security and privacy risks in the smart home. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 255–267. Springer, 2016.

[211] J. A. Obar and A. Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

[212] E. Okoyomon, N. Samarin, P. Wijesekera, A. Elazari Bar On, N. Vallina-Rodriguez, I. Reyes, Á. Feal, S. Egelman, et al. On the ridiculousness of notice and consent: Contradictions in app privacy policies. In *Workshop on Technology and Consumer Protection (ConPro 2019), in conjunction with the 39th IEEE Symposium on Security and Privacy*, 2019.

[213] E. Oriwoh and M. Conrad. Presence detection from smart home motion sensor datasets: a model. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pages 1249–1255. Springer, 2016.

[214] L. Pacheco, E. Alchieri, and P. Solis. Architecture for privacy in cloud of things. In *ICEIS (2)*, pages 487–494, 2017.

[215] A. Pakrashi, D. Greene, and B. MacNamee. Benchmarking multi-label classification algorithms. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), Dublin, Ireland, 20-21 September 2016*. CEUR Workshop Proceedings, 2016.

[216] Pallets. Flask. `https://flask.palletsprojects.com/en/2.3.x/`, 2023. [Online; Accessed April 2023].

[217] S. B. Pan. Get to know me: Protecting privacy and autonomy under big data's penetrating gaze. *Harv. JL & Tech.*, 30:239, 2016.

Alexie Dini

[218] P. Pappachan, M. Degeling, R. Yus, A. Das, S. Bhagavatula, W. Melicher, P. E. Naeini, S. Zhang, L. Bauer, A. Kobsa, et al. Towards privacy-aware smart buildings: Capturing, communicating, and enforcing privacy policies and preferences. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 193–198. IEEE, 2017.

[219] A. Parate. Designing efficient and accurate behavior-aware mobile systems. 2014.

[220] J. Park, K. Jang, and S.-B. Yang. Deep neural networks for activity recognition with multi-sensor data in a smart home. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, pages 155–160. IEEE, 2018.

[221] R. B. Parker. A definition of privacy. *Rutgers L. Rev.*, 27:275, 1973.

[222] J. Passos, S. I. Lopes, F. M. Clemente, P. M. Moreira, M. Rico-González, P. Bezerra, and L. P. Rodrigues. Wearables and internet of things (iot) technologies for fitness assessment: a systematic review. *Sensors*, 21(16):5418, 2021.

[223] S. T. Peek, E. J. Wouters, J. Van Hoof, K. G. Luijkx, H. R. Boeije, and H. J. Vrijhoef. Factors influencing acceptance of technology for aging in place: a systematic review. *International journal of medical informatics*, 83(4):235–248, 2014.

[224] S. R. Peppet. Regulating the internet of things: first steps toward managing discrimination, privacy, security and consent. *Tex. L. Rev.*, 93:85, 2014.

[225] C. Perera, C. McCormick, A. K. Bandara, B. A. Price, and B. Nuseibeh. Privacy-by-design framework for assessing internet of things applications and platforms. In *Proceedings of the 6th International Conference on the Internet of Things*, IoT'16, pages 83–92, New York, NY, USA, 2016. ACM.

[226] A. J. Perez, S. Zeadally, and J. Cochran. A review and an empirical analysis of privacy policy and notices for consumer internet of things. *Security and Privacy*, 1(3):e15, 2018.

[227] J. Pierce. Smart home security cameras and shifting lines of creepiness: A design-led inquiry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[228] E. Politou, E. Alepis, and C. Patsakis. Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of Cybersecurity*, 4(1):tyy001, 2018.

[229] A. Prince. Location as health. *Houston Journal of Health Law and Policy, Forthcoming, U Iowa Legal Studies Research Paper*, (2021-06), 2021.

[230] I. Psychoula, L. Chen, and O. Amft. Privacy risk awareness in wearables and the internet of things. *IEEE Pervasive Computing*, 19(3):60–66, 2020.

[231] I. Psychoula, L. Chen, and F. Chen. Privacy modelling and management for assisted living within smart homes. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE, 2017.

[232] V. Puri, P. Kaur, and S. Sachdeva. Data anonymization for privacy protection in fog-enhanced smart homes. In *2020 6th International Conference on Signal Processing and Communication (ICSC)*, pages 201–205. IEEE, 2020.

Alexia Dini

[233] A. Qamar, T. Javed, and M. O. Beg. Detecting compliance of privacy policies with data protection laws. *arXiv preprint arXiv:2102.12362*, 2021.

[234] J. Qi, P. Yang, M. Hanneghan, S. Tang, and B. Zhou. A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors. *IEEE Internet of Things Journal*, 6(2):1384–1393, 2018.

[235] M. Rahmany, A. M. Zin, and E. A. Sundararajan. Comparing tools provided by python and r for exploratory data analysis. *IJISCS (International Journal of Information System and Computer Science)*, 4(3):131–142, 2020.

[236] J. Ramesh, R. Aburukba, and A. Sagahyroon. A remote healthcare monitoring framework for diabetes prediction using machine learning. *Healthcare Technology Letters*, 8(3):45–57, 2021.

[237] K. Rantos, G. Drosatos, K. Demertzis, C. Ilioudis, A. Papanikolaou, and A. Kritsas. Advocate: a consent management platform for personal data processing in the iot using blockchain technology. In *Innovative Security Solutions for Information Technology and Communications: 11th International Conference, SecITC 2018, Bucharest, Romania, November 8–9, 2018, Revised Selected Papers 11*, pages 300–313. Springer, 2019.

[238] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, pages 254–269. Springer, 2009.

[239] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine learning*, 85:333–359, 2011.

[240] T. Reichherzer, M. Timm, N. Earley, N. Reyes, and V. Kumar. Using machine learning techniques to track individuals & their fitness activities. In *CATA 2017*, pages 119–124. ISCA, 2017.

[241] D. Reinhardt, J. Borchard, and J. Hurtienne. Visual interactive privacy policy: The better choice? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[242] L. Richardson. beautifulsoup4 4.11.1. `https://pypi.org/project/beautifulsoup4/`, 2023. [Online; Accessed April 2023].

[243] A. Rossi and M. Palmirani. Dapis: a data protection icon set to improve information transparency under the gdpr. *Knowledge of the Law in the Big Data Age. Frontiers*, 252:181–195, 2019.

[244] G. Sagirlar, B. Carminati, and E. Ferrari. Decentralizing privacy enforcement for internet of things smart objects. *Computer Networks*, 143:112–125, 2018.

[245] P. F. Saint-Maurice, R. P. Troiano, D. R. Bassett, B. I. Graubard, S. A. Carlson, E. J. Shiroma, J. E. Fulton, and C. E. Matthews. Association of daily step count and step intensity with mortality among us adults. *Jama*, 323(12):1151–1160, 2020.

[246] I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

Alexia Dini

[247] M. Sarstedt and E. Mooi. Descriptive statistics. In *A Concise Guide to Market Research*, pages 91–150. Springer, 2019.

[248] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, S. Taheri, et al. Sleep quality prediction from wearable data using deep learning. *JMIR mHealth and uHealth*, 4(4):e6562, 2016.

[249] S. Schiffner, B. Berendt, T. Siil, M. Degeling, R. Riemann, F. Schaub, K. Wuyts, M. Attoresi, S. Gürses, A. Klabunde, et al. Towards a roadmap for privacy technologies and the general data protection regulation: A transatlantic initiative. In *Annual Privacy Forum*, pages 24–42. Springer, 2018.

[250] M. Schwarzkopf, E. Kohler, M. Frans Kaashoek, and R. Morris. Position: Gdpr compliance by construction. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 39–53. Springer, 2019.

[251] R. Scitovski and K. Sabo. Dbscan-like clustering method for various data densities. *Pattern Analysis and Applications*, pages 1–14, 2019.

[252] M. Seliem, K. Elgazzar, and K. Khalil. Towards privacy preserving iot environments: a survey. *Wireless Communications and Mobile Computing*, 2018, 2018.

[253] P. C. Sen, M. Hajra, and M. Ghosh. Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, pages 99–111. Springer, 2020.

[254] J. Seo, K. Kim, M. Park, M. Park, and K. Lee. An analysis of economic impact on iot industry under gdpr. *Mobile Information Systems*, 2018, 2018.

[255] F. Shao, R. Cheng, and F. Zhang. A full privacy-preserving scheme for location-based services. In *Information and Communication Technology-EurAsia Conference*, pages 596–601. Springer, 2014.

[256] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy. Chatgpt and other large language models are double-edged swords, 2023.

[257] G. Shin, M. H. Jarrahi, Y. Fei, A. Karami, N. Gafinowitz, A. Byun, and X. Lu. Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. *Journal of biomedical informatics*, 93:103153, 2019.

[258] P. Shrestha and N. Saxena. An offensive and defensive exposition of wearable computing. *ACM Comput. Surv.*, 50(6):92:1–92:39, Nov. 2017.

[259] A. Shuhaiber and I. Mashal. Understanding users acceptance of smart homes. *Technology in Society*, 58:101110, 2019.

[260] T. Sigmund. Attention paid to privacy policy statements. *Information*, 12(4):144, 2021.

[261] A. P. Singh, V. Jain, S. Chaudhari, F. A. Kraemer, S. Werner, and V. Garg. Machine learning-based occupancy estimation using multivariate sensor nodes. In *2018 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2018.

Alexia Dini

[262] D. Singh, E. Merdivan, I. Psychoula, J. Kropf, S. Hanke, M. Geist, and A. Holzinger. Human activity recognition using recurrent neural networks. In *Machine Learning and Knowledge Extraction: First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio, Italy, August 29–September 1, 2017, Proceedings 1*, pages 267–274. Springer, 2017.

[263] A. Siraj et al. Avoiding occupancy detection from smart meter using adversarial machine learning. *arXiv preprint arXiv:2010.12640*, 2020.

[264] A. Skiljic. Health Inferences. `https://iapp.org/news/a/the-status-quo-of-health-data-inferences/`, 2021. [Online; Accessed April 2023].

[265] SleepFoundation. Sleep Foundation. `https://www.sleepfoundation.org/`, 2023. [Online; Accessed April 2023].

[266] M. Sobolewski, J. Mazur, and M. Paliński. Gdpr: A step towards a user-centric internet? *Intereconomics*, 52(4):207–213, 2017.

[267] D. J. Solove. Understanding privacy. 2008.

[268] I. Spectrum. The Consumer Electronics Hall of Fame: Fitbit. `https://spectrum.ieee.org/the-consumer-electronics-hall-of-fame-fitbit#toggle-gdpr`, 2019. [Online; Accessed April 2023].

[269] E. Spyromitros, G. Tsoumakas, and I. Vlahavas. An empirical study of lazy multilabel classification algorithms. In *Artificial Intelligence: Theories, Models and Applications: 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2-4, 2008. Proceedings 5*, pages 401–406. Springer, 2008.

[270] Statista. Smart Home - Worldwide. `https://www.statista.com/outlook/dmo/smart-home/worldwide/`, 2022. [Online; Accessed April 2023].

[271] G. Suciu, A. Vulpe, S. Halunga, O. Fratu, G. Todoran, and V. Suciu. Smart cities built on resilient cloud computing and secure internet of things. In *2013 19th international conference on control systems and computer science*, pages 513–518. IEEE, 2013.

[272] L. Sweeney, A. Abu, and J. Winn. Identifying participants in the personal genome project by name (a re-identification experiment). *arXiv preprint arXiv:1304.7605*, 2013.

[273] M. D. Szabó. Kísérlet a privacy fogalmának meghatározására a magyar jogrendszer fogalmaival. *Információs Társadalom: társadalomtudományi folyóirat*, 5(2), 2005.

[274] E. Szczekocka, J. Gromada, A. Filipowska, P. Jankowiak, P. Kałuzny, A. Brun, J. M. Portugal, and J. Staiano. Managing personal information: a telco perspective. *Proceedings of the 19th international innovations in clouds, internet and networks (ICIN)*, pages 1–8, 2016.

[275] Q. Tang. *Automated detection of puffing and smoking with wrist accelerometers*. Northeastern University, 2014.

[276] H. Tao and W. Peiran. Preference-based privacy protection mechanism for the internet of things. In *2010 Third International Symposium on Information Science and Engineering*, pages 531–534. IEEE, 2010.

Alexia Dini

[277] J. Tao and X. Fang. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7:1–26, 2020.

[278] N. Team. Natural Language Toolkit (NLTK). `https://pypi.org/project/nltk/`, 2023. [Online; Accessed April 2023].

[279] S. Tedesco, M. Sica, A. Ancillao, S. Timmons, J. Barton, and B. OFlynn. Accuracy of consumer-level and research-grade activity trackers in ambulatory settings in older adults. *PloS one*, 14(5):e0216891, 2019.

[280] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna. Privacyguide: towards an implementation of the eu gdpr on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 15–21, 2018.

[281] W. B. Tesfay and J. Serna-Olvera. Towards user-centered privacy risk detection and quantification framework. In *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5. IEEE, 2016.

[282] P. K. Thakkar, S. He, S. Xu, D. Y. Huang, and Y. Yao. it would probably turn into a social faux-pas: Users and bystanders preferences of privacy awareness mechanisms in smart homes. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[283] E. Thomaz, I. Essa, and G. D. Abowd. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1029–1040, 2015.

[284] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[285] D. Torre, S. Abualhaija, M. Sabetzadeh, L. Briand, K. Baetens, P. Goes, and S. Forastier. An ai-assisted approach for checking the completeness of privacy policies against gdpr. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 136–146. IEEE, 2020.

[286] I. Torre, G. Adorni, F. Koceva, and O. Sanchez. Preventing disclosure of personal data in iot networks. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 389–396. IEEE, 2016.

[287] I. Torre, F. Koceva, O. R. Sanchez, and G. Adorni. Fitness trackers and wearable devices: how to prevent inference risks? In *Proceedings of the 11th EAI International Conference on Body Area Networks*, pages 125–131. ICST (Institute for Computer Sciences, Social-Informatics and , 2016.

[288] I. Torre, F. Koceva, O. R. Sanchez, and G. Adorni. A framework for personal data protection in the iot. In *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 384–391. IEEE, 2016.

[289] I. Torre, O. R. Sanchez, F. Koceva, and G. Adorni. Supporting users to take informed decisions on privacy settings of personal devices. *Personal and Ubiquitous Computing*, 22(2):345–364, 2018.

Alexia Dini

[290] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Int. J. Data Warehous. Min.*, 3(3):1–13, 2007.

[291] G. Tsoumakas, I. Katakis, and I. Vlahavas. A review of multi-label classification methods. In *Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006)*, pages 99–109, 2006.

[292] C. Tudor-Locke and D. R. Bassett. How many steps/day are enough? *Sports medicine*, 34(1):1–8, 2004.

[293] T. S. Tullis and J. N. Stetson. A comparison of questionnaires for assessing website usability. In *Usability professional association conference*, volume 1, pages 1–12. Minneapolis, USA, 2004.

[294] A. Ukil, S. Bandyopadhyay, J. Joseph, V. Banahatti, and S. Lodha. Negotiation-based privacy preservation scheme in internet of things platform. In *Proceedings of the First International Conference on Security of Internet of Things*, pages 75–84. ACM, 2012.

[295] A. Ukil, S. Bandyopadhyay, and A. Pal. Privacy for iot: Involuntary privacy enablement for smart energy systems. In *2015 IEEE International Conference on Communications (ICC)*, pages 536–541. IEEE, 2015.

[296] I. Ullah and M. A. Shah. A novel model for preserving location privacy in internet of things. In *2016 22nd International conference on automation and computing (ICAC)*, pages 542–547. IEEE, 2016.

[297] I. Ullah, M. A. Shah, A. Wahid, A. Mehmood, and H. Song. Esot: a new privacy model for preserving location privacy in internet of things. *Telecommunication Systems*, 67(4):553–575, 2018.

[298] L. S. Vailshery. Iot connected devices worldwide 2030. `https://www.statista.com/statistics/802690/worldwide-connected-devices-by-access-technology/`, 2021. [Online; Accessed February 2023].

[299] P. Valdez. Focus: Attention science: Circadian rhythms in attention. *The Yale journal of biology and medicine*, 92(1):81, 2019.

[300] E. Vanezi, G. Zampa, C. Mettouris, A. Yeratziotis, and G. A. Papadopoulos. Complicy: Evaluating the gdpr alignment of privacy policies-a study on web platforms. In *International Conference on Research Challenges in Information Science*, pages 152–168. Springer, 2021.

[301] L. Vegh. A survey of privacy and security issues for the internet of things in the gdpr era. In *2018 International Conference on Communications (COMM)*, pages 453–458. IEEE, 2018.

[302] L. Velykoivanenko, K. S. Niksirat, N. Zufferey, M. Humbert, K. Huguenin, and M. Cherubini. Are those steps worth your privacy? fitness-tracker users' perceptions of privacy and utility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–41, 2021.

[303] K. Vemou, M. Karyda, and S. Kokolakis. Directions for raising privacy awareness in sns platforms. In *Proceedings of the 18th Panhellenic conference on informatics*, pages 1–6, 2014.

Alexia Dini

[304] J. Vitak, Y. Liao, P. Kumar, M. Zimmer, and K. Kritikos. Privacy attitudes and data valuation among fitness tracker users. In *International Conference on Information*, pages 229–239. Springer, 2018.

[305] I. Vuori. Physical inactivity is a cause and physical activity is a remedy for major public health problems. *Kinesiology*, 36(2):123–153, 2004.

[306] S. Wachter. Ethical and normative challenges of identification in the internet of things. 2018.

[307] S. Wachter. Gdpr and the internet of things: guidelines to protect users identity and privacy. *Tillgänglig online: https://papers. ssrn. com/sol3/papers. cfm*, 2018.

[308] S. Wachter. Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the gdpr. *Computer law & security review*, 34(3):436–449, 2018.

[309] S. Wachter and B. Mittelstadt. A right to reasonable inferences: Re-thinking data protection law in the age of big data and ai. *Colum. Bus. L. Rev.*, page 494, 2019.

[310] J. Wang, N. Spicher, J. M. Warnecke, M. Haghi, J. Schwartze, and T. M. Deserno. Unobtrusive health monitoring in private spaces: The smart home. *Sensors*, 21(3):864, 2021.

[311] D. E. Webster, M. Tummalacherla, M. Higgins, D. Wing, E. Ashley, V. E. Kelly, M. V. McConnell, E. D. Muse, J. E. Olgin, L. M. Mangravite, et al. Smartphone-based vo2max measurement with heart snapshot in clinical and real-world settings with a diverse population: Validation study. *JMIR mHealth and uHealth*, 9(6):e26006, 2021.

[312] A. F. Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[313] A. F. Westin. Social and political dimensions of privacy. *Journal of social issues*, 59(2):431–453, 2003.

[314] Z. Whittaker. How Strava's "anonymized" fitness tracking data spilled government secrets. `https://www.zdnet.com/article/strava-anonymized-fitness-tracking-data-government-opsec/`, 2018. [[Online; Accessed February 2023].

[315] C. I. Wickramasinghe and D. Reinhardt. A user-centric privacy-preserving approach to control data collection, storage, and disclosure in own smart home environments. In *Mobile and Ubiquitous Systems: Computing, Networking and Services: 18th EAI International Conference, MobiQuitous 2021, Virtual Event, November 8-11, 2021, Proceedings*, pages 190–206. Springer, 2022.

[316] C. Wilson, S. Lina, V. Stankovic, J. Liao, M. Coleman, R. Hauxwell-Baldwin, T. Kane, S. Firth, and T. Hassan. Identifying the time profile of everyday activities in the home using smart meter data. 2015.

[317] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, 2016.

Alexia Dini

[318] S. Winkler and S. Zeadally. Privacy policy analysis of popular web platforms. *IEEE technology and society magazine*, 35(2):75–85, 2016.

[319] H. Wu, B. P. Knijnenburg, and A. Kobsa. Improving the prediction of users' disclosure behavior by making them disclose more predictably? In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[320] Q. Wu, K. Sum, and D. Nathan-Roberts. How fitness trackers facilitate health behavior change. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 1068–1072. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[321] S. Xu, G. Yang, Y. Mu, and X. Liu. A secure iot cloud storage system with fine-grained access control and decryption key exposure resistance. *Future Generation Computer Systems*, 97:284–294, 2019.

[322] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, pages 261–270, 2009.

[323] T. Yan, Y. Lu, and N. Zhang. Privacy disclosure from wearable devices. In *Proceedings of the 2015 Workshop on Privacy-Aware Mobile Computing*, pages 13–18. ACM, 2015.

[324] L. Yang, H. Deng, and X. Dang. Preference preserved privacy protection scheme for smart home network system based on information hiding. *IEEE Access*, 8:40767–40776, 2020.

[325] L. Yang, H. Deng, R. P. Liu, P. Wang, X. Dang, Y. Y. Tang, and X. Li. Smart home privacy protection based on the improved lsb information hiding. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(12):2160005, 2021.

[326] L. Yang, K. Ting, and M. B. Srivastava. Inferring occupancy from opportunistically available sensor data. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 60–68. IEEE, 2014.

[327] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao. A survey on security and privacy issues in internet-of-things. *IEEE Internet of Things Journal*, 4(5):1250–1258, 2017.

[328] Y. Yang, X. Zheng, W. Guo, X. Liu, and V. Chang. Privacy-preserving smart iot-based healthcare big data storage and self-adaptive access control system. *Information Sciences*, 479:567–592, 2019.

[329] Z. Yang. The Backend Architecture of a Python Web App. `https://medium.com/techtofreedom/backend-architecture-of-a-python-web-application-7af256ee004c`, 2020. [Online; Accessed April 2023].

[330] Y. Yao, L. Song, and J. Ye. Motion-to-bmi: Using motion sensors to predict the body mass index of smartphone users. *Sensors*, 20(4):1134, 2020.

[331] E. K. Yapp, X. Li, W. F. Lu, and P. S. Tan. Comparison of base classifiers for multi-label learning. *Neurocomputing*, 394:51–60, 2020.

Alexia Dini

[332] A. Yassine, S. Singh, and A. Alamri. Mining human activity patterns from smart home big data for health care applications. *IEEE Access*, 5:13131–13141, 2017.

[333] M. Yedla, S. R. Pathakota, and T. Srinivasa. Enhancing k-means clustering algorithm with improved initial center. *International Journal of computer science and information technologies*, 1(2):121–125, 2010.

[334] Z. Yu, Q. Wang, Y. Fan, H. Dai, and M. Qiu. An improved classifier chain algorithm for multi-label classification of big data analysis. In *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, pages 1298–1301. IEEE, 2015.

[335] E. Zeng, S. Mare, and F. Roesner. End user security and privacy concerns with smart homes. In *Symposium on Usable Privacy and Security (SOUPS)*, volume 220, 2017.

[336] H. Zhang, Z. Fu, and K.-I. Shu. Recognizing ping-pong motions using inertial data based on machine learning classification algorithms. *IEEE Access*, 7:167055–167064, 2019.

[337] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12:191–202, 2018.

[338] M.-L. Zhang and Z.-H. Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE international conference on granular computing*, volume 2, pages 718–721. IEEE, 2005.

[339] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[340] S. Zhang, J. Rong, and B. Wang. A privacy protection scheme of smart meter for decentralized smart home environment based on consortium blockchain. *International Journal of Electrical Power & Energy Systems*, 121:106140, 2020.

[341] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster. User perceptions of smart home iot privacy. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20, 2018.

[342] X. Zheng, Z. Cai, and Y. Li. Data linkage in smart internet of things systems: a consideration from a privacy perspective. *IEEE Communications Magazine*, 56(9):55–61, 2018.

[343] M. Zimmer, P. Kumar, J. Vitak, Y. Liao, and K. Chamberlain Kritikos. theres nothing really they can do with this information: unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society*, 23(7):1020–1037, 2020.

[344] V. Zimmermann, M. Bennighof, M. Edel, O. Hofmann, J. Jung, and M. von Wick. home, smart home–exploring end users mental models of smart homes. *Mensch und Computer 2018-Workshopband*, 2018.

Alexia Dini

Chapter A

List of GDPR user right Terms

A.1 The Right to Information Terms File

- categories of personal
- categories of recipients
- collect your personal
- collection of data
- collection of information
- collection of personal data
- data collect
- data do we collect
- data retention
- data sharing
- data transfer
- data we collect
- disclose your personal
- disclosing your information
- disclosure of data
- disclosure of personal
- disclosure of your information
- disclosure of your personal data
- disclosure to third
- disclosures of your personal data
- do with your information
- duration of the processing
- handle your personal information
- how is the information collected and used
- how long do we keep your information
- how we collect
- how we share
- how your personal data is collected
- how your personal information is collected
- information collect
- information disclosure
- information do we collect
- information is shared
- information retention
- information shar
- information that we share
- information use
- information we collect

Alexia Dini

- information we process
- information we share
- information we transfer
- international transfer
- legal basis for
- opt out
- opt-out
- personal data is collected
- personal data that is processed
- personal information be transferred
- process personal information
- process your personal data
- providing your personal data to others
- purpose and basis for processing
- purpose of processing
- purposes for collect
- purposes of our data collection
- purposes of processing
- purposes of the processing
- recipients of personal data
- retain your information
- retention of data
- retention of personal
- retention of your personal
- retention period
- retention policy
- retention time
- right to be informed
- right to information
- right to know
- share personal
- share your data
- share your information
- share your personal
- sharing information
- sharing of information
- sharing personal information
- sharing your personal information
- transfer of data
- transfer of personal data
- transfer your personal data
- transfers of personal data
- types of data
- types of information
- types of personal
- use and disclosure of information
- use information
- use of data
- use of information
- use of personal

Alexia Dini

- use of your information
- use personal data
- use the information
- use your data
- use your information
- use your personal
- uses of personal
- what information is collected
- who we give your information
- withdraw consent
- withdraw such consent
- withdraw your consent
- withdrawal of consent
- withdrawal of your consent

A.2 The Right of Access Terms file

- access personal data
- access personal information
- access right
- access their data
- access to personal data
- access your data
- access your personal data
- access your personal information
- change and delete your personal
- objecting to data use
- obtain access
- request a copy of your information
- request a copy of your personal data
- request a copy of your personal information
- request access
- request access to a copy of your personal data
- revoke the access
- right of access
- right to access
- right to access, correct, update and delete personal data
- right to complaint
- right to file a complaint
- right to file complaint
- right to information
- right to lodge a complaint
- right to lodge complaints
- right to obtain a copy
- right to request and receive information
- rights to access

Alexia Dini

A.3 The Right to Rectification Terms File

- correct any inaccuracies
- request to review
- change and delete your personal information
- correct your information
- rectification of your personal data
- rectify
- rectify or delete your personal data
- rectify your data
- request rectification
- request that we correct
- request that we update
- request the rectification
- right of correction
- right of rectification
- right to access, correct, update and delete personal data
- right to amend or update
- right to complete incomplete personal data
- right to correct
- right to correct and update
- right to correction
- right to delete or modify any personal information
- right to have incomplete personal data
- right to rectification
- right to request correction
- right to request Proper rectification
- right to request rectification
- right to request that we rectify or correct
- right to request the correction
- right to request update
- right to update
- seek rectification
- update or correct your information
- update your information

Alexia Dini

A.4 The Right to Erasure Terms File

- ask to erase
- ask us to erase
- delete of your account
- delete personal information
- delete your personal information
- erase any personal data
- erase personal data
- erase the personal data
- erase your information
- erase your personal data
- obtain the erasure of their data

- rectify or delete your personal data
- request deletion
- request erasure
- request that we delete
- request that we erase
- request that your personal data be deleted
- request the deletion
- request the erasure
- request to delete your data
- right of erasure
- right to access, correct, update and delete personal data
- right to be forgotten
- right to delete
- right to erasure
- right to request deletion
- right to request that we delete your data
- right to request the deletion
- to erase your data

A.5 The Right to Restriction of Processing Terms File

- object to our processing of your data
- object to, or limit or restrict, use of data
- processing be restrictedă
- request restriction of personal data processing
- request restrictions
- request the restriction of data use
- request the restriction of personal data use
- request the restriction of their use
- restrict or limit processing
- restrict or limit the processing
- restrict our processing
- restrict our uses of your personal information
- restrict the personal information
- restrict the processing
- restrict your data
- restriction of processing
- right of data subjects to be informed about the restriction
- right of restriction
- right to demand processing restrictions
- right to object to certain types of processing
- right to propose other restriction
- right to restrict
- right to restriction
- restrict processing

A.6 The Right to Data Portability Terms File

- ask for a copy of your personal data

Alexia Dini

- request portability
- request the transfer data
- request the transfer of your personal data
- request the transfer personal data
-
- request the transfer your personal data
- right of data portability
- right of portability
- right to data portability
- right to obtain your personal information copy
- right to portability
- right to receive a copy of your personal information
- right to receive a subset of the personal data
- right to receive personal data
- right to receive the personal data
- right to receive your personal data
- right to the portability of your data
- right to transmit data
- right to transmit my data
- right to transmit personal data
- right to transmit those data
- the right to transmit those data
- transmit your data
- transmit your personal data

A.7 The Right to Object Terms File

- object processing of personal data
- object to our processing
- object to processing
- object to the further processing
- object to the processing of his/her data
- object to the processing
- object to your personal information being processed
- object to, or limit or restrict, use of data
- processing objection
- further processing
- right to object
- right to object at any time to processing
- right to object at any time to processing of personal data
- right to object to certain types of processing
- right to object to processing

A.8 The Right to Avoid Automated Decision-Making Terms File

- automated decision-making, including profiling
- object to a decision based solely on automated processing, including profiling

Alexia Dini

- object to automated decision-making
- objecting to automated decision making and profiling
- right not to be subject to a decision based solely on automated processing
- right not to be subject to a decision which is based solely on automated processing
- right to object to automated decision making
- right to refuse to be subjected to automated decision making, including profiling
- rights related to automated decision making including profiling
- automated decision-making

Alexia Dini