TOWARDS A DESIGN ECOSYSTEM FOR A PERSONAL DIGITAL TWIN FOR WELL-BEING

by

ROBERTO ALEJANDRO MARTINEZ VELAZQUEZ

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

uOttawa

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

This thesis focuses on Personal Digital Twins (PDTs) for well-being and their transformative impact on healthcare. It proposes an integration of technologies such as virtual reality, artificial intelligence (AI), machine learning (ML), and data analytics, which opens new avenues for research in eHealth and Personal Digital Twins for well-being. The thesis presents the design of an ecosystem for implementing PDTs for healthcare, which extends beyond the traditional boundaries of telemedicine. This ecosystem is scalable and interoperable, capable of handling real-time health data from various sources, including electronic health records, wearable sensors, and patient-generated data.

The core of the thesis explores the challenges of designing a digital twin that not only replicates a patient's physical state but also interacts dynamically with health data to offer personalized care. The study emphasizes the role of machine learning (ML) in processing vast amounts of health data, enabling predictive health insights and enhancing the decision-making process in clinical settings. This integration also opens the door for practical applications that promise to revolutionize patient monitoring, diagnostics, and treatment planning, moving towards a more proactive and preventive healthcare model. The thesis presents the implementations of two PDTs proof of concepts, namely Cardio Twin and COVIDMe.

However, deploying PDTs for health is not without challenges; this thesis discusses the potential challenges in adopting this technology. It sets a foundational blueprint for the future exploration of personal digital twins for health. It paves the way for innovative research lines, advocating for a collaborative approach that involves researchers, healthcare professionals, engineers, and policymakers.

Acknowledgements

I thank Dr. Abdulmotaleb El Saddik for his continuous support throughout the program; succeeding in this program was only possible with his guidance and vast experience. His friendship and positive attitude inspired me to do my best in the program. I will remember the lab meetings; as tired as I could have felt, I would always come out with renewed energy and drive to continue pushing my limits and “enhancing my knowledge.” The Ph.D. is about research and the drive to generate knowledge, but it is also about life and people; he understands it and is always there for his students.

I want to thank my friends and colleagues in the MCR Lab. I was so lucky to be in a place full of capable and intelligent people who were always willing to share their knowledge and guide me.

Thank you to my mother, “Edi,” who has supported and encouraged me since my first day of school, back when I was a child. The funny thing about parents is that sometimes they tell small lies. She told me I was intelligent, and I believed her—or maybe she truly has always believed in me. In any case, her encouragement has continuously pushed me to give my best and succeed.

Thanks to the CONACyT for funding part of my expenses as a Ph.D. student.

Yo solo podría lograr poco, con ustedes puedo lograrlo todo. Para Ana, Lucia y Roberto. Cuando estuve cansado, me dieron fuerza; cuando había incertidumbre, me mostraron el camino; en el desánimo, fueron mi inspiración; gracias a ustedes nunca me sentí solo y siempre pude refugiarme en su amor.

Table of Contents

List of Tables

List of Algorithms

List of Figures

Chapter 1

1. Introduction

The advent of Personal Digital Twins (PDTs) represents a significant leap forward in healthcare and well-being. Although a detailed definition of PDTs is presented in Chapter 2, an interim definition is needed for the current chapter; for the moment, let a PDT be understood as a detailed digital copy of an individual. By using machine learning (ML) to analyze health and behavioural data, PDTs offer a pathway to insights and predictions that can be leveraged to improve individual well-being. This thesis explores the design of PDTs, focusing on their potential to drive advancements in health monitoring, diagnosis, and treatment. This research work aims to contribute to the field with a shared design ecosystem for designing PDTs for well-being.

The concept of Digital Twins, initially rooted in industrial applications, has evolved to encompass the personal dimension, giving rise to Personal Digital Twins. PDTs are virtual replicas of individuals, meticulously crafted and updated with a wealth of data ranging from physiological signals (i.e. heart rate, body temperature, etc.) to behavioural patterns. These digital counterparts harness the power of machine learning to process and analyze data, drawing insights and predictions that are then used to improve the individual’s well-being. The transition from industrial applications to healthcare has been seamless yet impactful, with PDTs now at the vanguard of individual health monitoring and disease management.

Adopting PDTs in healthcare signifies a paradigm shift towards more proactive and personalized medical care. Unlike traditional healthcare models that often adopt a reactive stance, PDTs enable continuous, dynamic monitoring and assessment of an individual's health. This is particularly beneficial in the management of chronic conditions such as ischemic heart disease and cancer, where early detection and ongoing management can dramatically influence outcomes. By integrating real-time health and behavioural data collection into the assessment, diagnosis and treatment processes, PDTs facilitate more effective healthcare delivery, underscoring the potential of digital health technologies to revolutionize patient care.

For this thesis, well-being is understood as a broader concept that covers but is not limited to physical health, psychological state, social relationships and how individuals relate to their environment. The World Health Organization (WHO) considers well-being an integral part of health, highlighting that health is a state of complete physical, mental and social well-being [1].

Extensive research on Personal Digital Twins (PDT) for well-being has been conducted in developing this thesis. This exploration has not only involved theoretical studies but also included the practical experience of constructing PDTs for well-being. Through this process, a series of guidelines have been formulated to inform the design of new PDTs for well-being. Rather than merely enumerating these guidelines, they have been organized into a design ecosystem dubbed Twin Me Well (TMW).

Designing personal digital twins for well-being requires a holistic approach encompassing various technical, ethical, and practical considerations. A PDT for well-being must support the seamless collection, transmission, and analysis of personal health and behavioural data, ensuring accuracy, privacy, and security at every juncture. Moreover, it must support the integration of machine learning algorithms that can adapt and evolve in response to new data, refining their output over time. The TMW design ecosystem facilitates the effective design of PDTs for well-being. It lays the foundation for their widespread adoption in research, thus resonating with the growing interest from the research community in leveraging PDTs for healthcare delivery.

The healthcare sector is increasingly adopting telemedicine as a viable option for healthcare delivery, which makes the potential of PDTs to transform the landscape of medical care undeniable. The PDTs offer a pathway to more personalized, proactive, and predictive healthcare tailored to everyone’s unique needs and conditions. The TMW framework contributes to realizing this potential. The promise of PDTs in enhancing well-being and managing diseases more effectively offers a glimpse into a future where technology and healthcare converge to improve the quality of life for individuals around the globe [2], [3].

1.1. Motivation

The gap between the digital and physical worlds continues to close thanks to the growing interest from the scientific community in Digital Twins. On the other hand, Personal Digital Twins (PDT) have also gained traction within the private and public sectors. These factors create novel avenues in eHealth, offering to transform clinical trials, public health, medical interventions and other aspects of healthcare and well-being. Yet, the transition from concept to implementation is hindered by significant technical hurdles. Implementing a PDT necessitates a robust design capable of handling the vast and heterogeneous data streams going into it while ensuring its availability and overall performance. The core challenge lies in the seamless fusion of diverse data sources: electronic health records, sensors collecting biological signals, specialized hardware collecting vitals, biometric sensors, environmental sensors and virtually any other type of sensor that has the potential to inform about the individual or its environment. The second major challenge is the need for sophisticated data processing techniques and the computational complexity and networking aspects of managing real-time data collection and transmission [4], [5], [6].

This thesis addresses these challenges by proposing a design ecosystem for personal digital twins for well-being while exploring technical implementation strategies that factor in ensuring scalability and interoperability; the design ecosystem has been named Twin Me Well. This research work also aims to bridge the gap between the theoretical potential of PDTs for well-being and their practical, real-world application in enhancing well-being and lay the foundation for scientific research on this topic.

Over the years, research efforts have explored the integration of Digital Twins (DT) and healthcare; services provided to individuals or communities for the promotion, maintenance, monitoring, and restoration of well-being and, consequently, health. However, no framework exists in the revised literature for this research that integrates these disparate efforts into a cohesive ecosystem.

The primary challenge is designing an ecosystem that facilitates the real-time rendering and analysis of health data, digital twin technology, and well-being innovations to foster a seamless, interactive, self-directed well-being management ecosystem.

1.2. Problem statement

In the wake of an increasingly digitized age, Personal Digital Twins have emerged as a frontier that impacts different sectors, industries, education, and health, among others. In recent years, the health sector has been experiencing a rapid evolution, with telemedicine being a focal point. Much of these efforts are being invested into more specific tasks of telemedicine, i.e., health monitoring, disease detection, and the overall mental well-being of individuals. Personal Digital Twins are transcendent to this transformation thanks to the potential of creating digital copies of people [5]. This thesis confronts the challenge of designing a real-time, responsive personal Digital Twin (DT) design ecosystem for well-being, named Twin Me Well (TMW). The design ecosystem may be used to inform the design of new PDTs capable of processing and analyzing health and environmental data, comprising self-reported symptoms, environmental data and physiological signals. The pivotal concern lies in the meticulous orchestration of data flow, from collection through local or cloud storage to intelligent feedback generation. The system must ensure data integrity, curation efficiency, and the accuracy of health feedback provided to the person, all while operating under real-time processing constraints. Additionally, the ecosystem must adeptly manage errors and exceptional flows.

1.3. Contributions

The design of a Personal Digital Twin framework for health and well-being. The contribution lies in creating a foundational Personal Digital Twin framework that caters to health and well-being.

This framework is a template for designing PDT systems that enhance individual health outcomes. Its significance is in offering a standardized approach that can be adapted for various health applications, promoting best practices, and accelerating innovation in personalized healthcare. A detailed list of contributions is next:

- The design of a Personal Digital Twin to assess the health state of the human heart in relation to the presence of ischemic heart diseases using real-time electrocardiogram signals. This contribution involves the design of a Personal Digital Twin specifically for monitoring the human heart's health, particularly in identifying ischemic heart diseases through the analysis of real-time electrocardiogram (ECG) signals. The relevance is the potential for early detection and ongoing management of heart conditions, contributing to improved patient outcomes and reducing healthcare burdens associated with heart diseases.
- The design of a classifier to detect abnormal Electrocardiogram signals in the context of ischemic heart diseases. Developing a classifier for detecting abnormalities in ECG signals related to ischemic heart diseases represents a significant stride in diagnostic tools. This contribution is important because it applies machine learning to enhance the accuracy and efficiency of diagnosing heart diseases, facilitating timely interventions and personalized treatment plans.
- The design of a Personal Digital Twin to assess the presence of COVID-19 infections based on self-reported symptoms via surveys. This contribution is the design of a Personal Digital Twin that assesses the likelihood of COVID-19 infection using data from self-reported symptoms via surveys. The relevance is twofold: it offers a non-invasive, user-friendly method for potential early detection of COVID-19, and it underscores the flexibility of PDTs to adapt to emergent health crises, demonstrating the utility of digital twins in the pandemic response.
- The design of a classifier capable of detecting COVID-19 from self-reported symptoms is a significant research contribution. It provides a tool for rapid screening that can complement traditional testing methods. This classifier is particularly relevant in situations where access to clinical testing is limited, enabling broader monitoring and potential early isolation to mitigate the spread of the virus.

1.4. Scholarly Achievements

- **Martinez-Velazquez, R.**, Gamez, R., & Saddik, A. el. (2019). Cardio Twin: A Digital Twin of the human heart running on the edge. 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 1–6. https://doi.org/10.1109/MeMeA.2019.8802162
- El Saddik, A., Badawi, H., **Martinez, R.**, Laamarti, F., Diaz, R., Bagaria, N., & Arteaga-Falconi, J. (2019). Dtwins: A Digital Twins Ecosystem For Health And Well-Being. IEEE COMSOC MMTC Communications -Frontiers, 14, 39–46.
- Bagaria, N., Laamarti, F., Badawi, H. F., Albraikan, A., **Martinez Velazquez, R. A.**, & el Saddik, A. (2020). Health 4.0: Digital Twins for Health and Well-Being. In Connected Health in Smart Cities (pp. 143–152). Springer International Publishing. https://doi.org/10.1007/978-3-030-27844-1_7
- **Martinez-Velazquez, R.**, Tobón V., D. P., Sanchez, A., el Saddik, A., & Petriu, E. (2021). A Machine Learning Approach as an Aid for Early COVID-19 Detection. In Sensors (Vol. 21, Issue 12). https://doi.org/10.3390/s21124202
- **Martinez-Velazquez, R.**, Ceballos, F., Sanchez, A., el Saddik, A., & Petriu, E. (2023). COVIDMe: a digital twin for COVID-19 self-assessment and detection. In Digital Twin for Healthcare (pp. 137–156). Elsevier. https://doi.org/10.1016/B978-0-32-399163-6.00012-3
- Faisal, M., **Martinez Velazquez, R. A.**, Laamarti, F., & el Saddik, A. (2023). Underactuated digital twins' robotic hands have tactile sensing capabilities for well-being. In Digital Twin for Healthcare (pp. 15–38). Elsevier. https://doi.org/10.1016/B978-0-32-399163-6.00007-X
- Wang, H., **Martinez-Velazquez, R.**, Dong, H., Saddik, A. el. (2024). Experimental Studies of Metaverse Streaming. IEEE Consumer Electronics Magazine, 1–11. https://doi.org/10.1109/MCE.2024.3364118

1.5. Thesis Organization

The remainder of this thesis is distributed in five more chapters. Chapter 2 presents an overview of relevant research work and a comprehensive conceptual framework for this thesis. Chapter 3 presents the design ecosystem for a Personal Digital Twin for health. Chapters 4 and 5 present the design of two PDTs for well-being, named CovidMe and Cardio Twin, respectively as a proof of concept for proposed design ecosystem. Finally, chapter 6 presents the conclusions and future work.

Chapter 2

2. Related work DT for Well-being

This chapter explores the convergence of Personal Digital Twins and eHealth. Digital Twins (DT) are relevant in industry applications and are rapidly becoming a frontier for healthcare and well-being delivery. Within this context, Personal Digital Twins (PDT), which are virtual replicas of people, are gaining relevance as they have the potential to create digital copies of a person's physical and mental state, which can be transmitted to a healthcare provider. This chapter provides a foundational understanding of these concepts and sets the stage for a deeper exploration of their integration. The chapter starts by introducing a holistic approach to healthcare, which sets the stage for a better understanding of how a PDT can improve the health of the PT. The chapter then introduces the relevant work in Personal Digital Twins assisting in healthcare and well-being.

2.1. Well-being: the big picture and a holistic approach to healthcare

The World Health Organization (WHO) defines health as a “state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” [7]. This definition has three base components or dimensions of health: physical, mental, and social well-being. Jason Raibley, who explains well-being as agential flourishing, offers one approach to understanding the concept of well-being.

The concept of well-being as agential flourishing is deeply rooted in the philosophical understanding of what constitutes a good life. According to the values-based approach presented in [8], well-being is determined not by an agent's desires or enjoyments but by their values. This perspective emphasizes that an agent, defined as a person who regulates its activities by referencing its values, flourishes when living according to them. The agential flourishing theory has practical implications in various fields, including healthcare. It has been suggested that this approach can be adapted for quality-of-life research, fitting the needs and aims of value assessment within the healthcare sector. This adaptation also influences debates about the fairness of healthcare allocations, indicating its broader impact on policy and ethical considerations. It offers a framework to understand well-being in situations of precariousness, which can be critical for policy-making and supporting individuals in their own environment and according to their socioeconomic situation. Human well-being or flourishing encompasses many states and outcomes, including mental and physical health, happiness, life satisfaction, meaning, purpose, character, virtue, and close social relationships. The empirical literature on this topic is vast, suggesting that flourishing is a multi-faceted concept that impacts various dimensions of an individual's life [8], [9].

Revisiting the WHO definition of health, which is defined as “a state of complete physical, mental and social well-being” [7], it can be argued that health is positively affected by procuring well-being. Hence, within the scope of this thesis, procuring our well-being ultimately contributes to our health. Furthermore, a state of happiness in a eudaemonic[1] way enhances the immune system and overall physical and mental health. Happiness is also a chemical process. The relations and dynamics of health, mental well-being and happiness are explored in psychoneuroimmunology. It is said that there is a correlation between eudaemonic happiness and a more robust immune system. This would mean that procuring happiness has as a consequence an increased protection against infections, thus helping the body present a more robust response against new threats and diseases [10], [11], [12], [13].

In conclusion, actively pursuing eudaemonic happiness and taking deliberate steps to maintain well-being can help improve physical health and elicit more favourable results from any medical treatments. A PDT for well-being would serve as a proactive tool to monitor health indicators, collect data from the PT, and encourage behaviours that support the broader and more holistic

---

[1] Eudaemonic happiness centers on living a life of purpose and meaning, where true contentment stems from personal growth, fulfilling one's potential, and aligning actions with deeply held values. It's a profound, enduring form of well-being that transcends momentary pleasures, focusing instead on achieving a sense of completeness and self-realization.

concept of health, as presented by the WHO. By integrating personalized health data, like biological signals and medical records, with a holistic understanding of well-being, a PDT could contribute to a more proactive healthcare model where individuals are empowered to take charge of their health and well-being, leading to improved physical health and more effective treatments.

The upcoming section will explore existing examples of PDTs and their influence on health and wellness.

2.2. Digital Twin

David Gelernter introduced Mirror Worlds, a collection of "software models of some chunk of reality, some piece of the real world going on outside your window." Gelernter's vision involves creating a software model of the real world through computer-assisted design. This model would receive a vast amount of data from the real world in real-time until it becomes an exact copy, creating a "mirror world." The mirror world would be so thorough and detailed that one could have a replica of a university, know the precise location of each student, faculty, and staff member, and assess the structural integrity of the buildings and every other little detail in between [14].

Michael Grieves introduced a Digital Twin (DT) in 2002. The DT is based on the original concept of “Mirror Worlds.” In this new framework, the real-world object is known as the Physical Twin (PT), i.e., the Digital Twin (DT) counterpart. An improvement over a Mirror World is that DT runs inside a Digital Twin Environment (DTE), and by leveraging statistical and mathematical models, the DTE can run simulations to predict outcomes of interactions between a PT and its environment. In this iteration, the DT was thought to mostly clone objects like buildings, vehicles, or production lines. The key to cloning the PTs in this framework is using sensors to collect data [15], [16]. When designing a PDT for health and well-being, the data can be extensive and derived from various sensors tracking a person's health and lifestyle. In the Grieves framework, in order to make sense of all this data, it would need to be funnelled into a DTE to create a digital replica, i.e., the DT of a person. In turn, the data can be leveraged to provide insights and forecast potential future states of the DT by applying Machine Learning[2] (ML).

An illustrative example would effectively demonstrate the operational aspects of the framework in a practical setting. A digital twin designed to monitor the human heart is a good design exercise. The goal is to create a Digital Twin that can track heart health, predict potential cardiac issues, and assist in managing conditions such as ischemic heart disease.

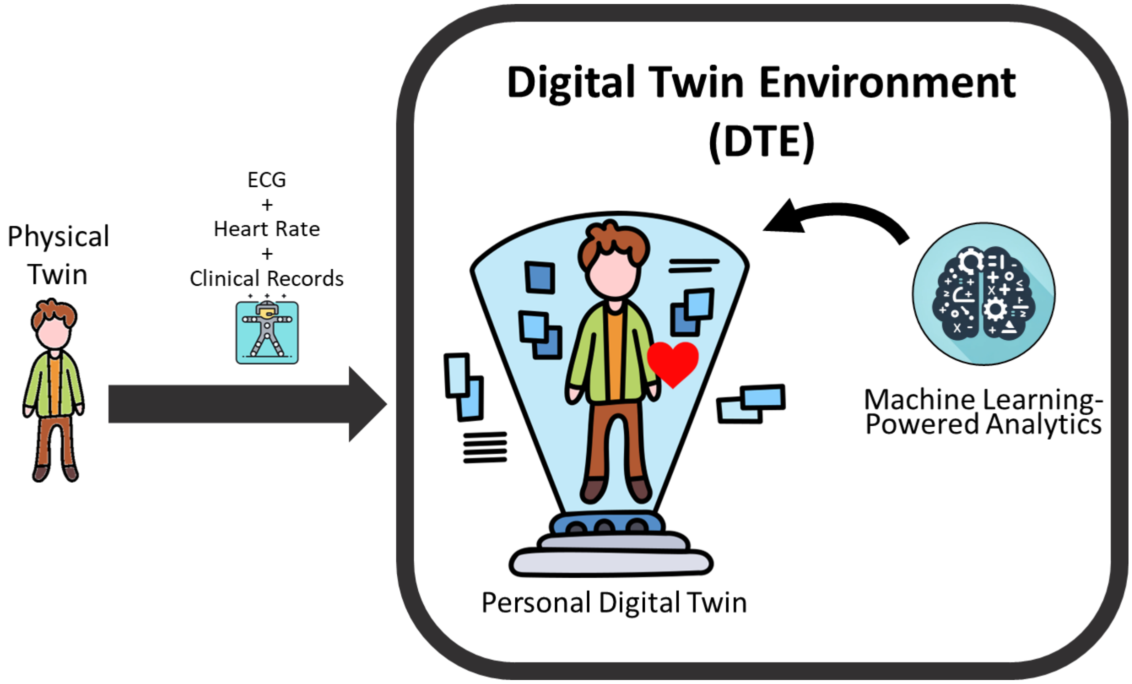


Figure 2-1. Digital Twin of the human heart, as it would be if it were designed according to the existing Digital Twin design paradigm. It provides a high-level overview of how data would flow through the system.

Figure 2-1 shows a representation of a DT of the human heart using the DT paradigm described above. The process begins with data collection from the person using a variety of hard and soft sensors. These might include hardware sensors that monitor heart rate, ECG, other biosignals, and data sources like clinical records or laboratory results. The collected data, which offers a

---

[2] Machine Learning (ML) is a field of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to perform a specific task without using explicit instructions, relying instead on patterns and inference. It involves computers learning from data provided, identifying patterns, and making decisions with minimal human intervention. ML applications range from email filtering and computer vision to self-driving cars, and it's a key technology in the development of intelligent systems.

comprehensive view of the individual's cardiac health, is then transmitted to the Digital Twin Environment (DTE).

Within the DTE, advanced Machine Learning (ML) algorithms process the data to create a digital clone of the person's heart. This Digital Twin continuously updates to reflect the person's current state and uses predictive analytics to forecast future heart health scenarios. Healthcare professionals can use these insights to make informed decisions about interventions, lifestyle recommendations, and treatment plans.

For example, if the Digital Twin indicates an increased cardiac event risk, preventative measures can be taken well in advance. This could involve adjusting medication, recommending dietary changes, or planning for surgical interventions if necessary. The predictive capabilities of the Digital Twin, powered by ML, also open the door for personalized healthcare strategies that adapt to changes in the patient's condition over time.

However, in the use case of PDT for the human heart, as in other similar use cases for PDT for well-being, the absence of a feedback mechanism directly involving the user could significantly impact the effectiveness of interventions. Without user feedback, there's a risk of a disconnect between the data-driven insights generated by the PDT and the individual's subjective experience and adherence to prescribed interventions. Personal feedback is essential for calibrating the Digital Twin's algorithms to the individual's unique responses and preferences. This oversight could lead to less tailored interventions and, potentially reduce patient engagement and adherence and, ultimately, have a negative effect on the success of health outcomes. Implementing a robust user feedback loop could enhance the personalization and effectiveness of the PDT, leading to more positive health interventions.

A DT is usually presented as a virtual copy of its physical counterpart, the Physical Twin (PT); however, such a description does not forcibly imply that a DT would always have to reflect the state of the PT in real time. A DT may also be described as a "set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level” [16]. The implication is that a DT can model behaviour from the PT by leveraging “virtual information constructs” (i.e., models) that help predict behaviour.

A Digital Twin Environment (DTE) maintains the data, processes, and models that render a Digital Twin.

The Digital Twin offers a distinct advantage over the "mirror worlds" paradigm. The DT paradigm doesn't just duplicate a system's current or past states but actively engages with the data to predict trends. In the case of a Digital Twin of the human heart, the DTE can analyze data from various sensors to understand the current health status and interrogate the model to predict potential ischemic events or other cardiac conditions before they occur.

This predictive ability of the DTE allows for pre-emptive healthcare interventions, personalized treatment plans, and continuous monitoring, which are crucial in managing chronic diseases and promoting well-being. For instance, by simulating different lifestyle scenarios or medical interventions within the DTE, healthcare providers can foresee the effects of certain behaviours or treatments on heart health. This can lead to more informed decisions that might prevent adverse events, improve patient outcomes, and reduce healthcare costs by avoiding unnecessary treatments or hospitalizations.

A PDT may leverage sensors to collect data from the PT and clone one or more attributes of a PT to clone different aspects of an actual person, from biological signals (i.e. ECG or heart rate) to parametric characteristics of an individual, such as body build, weight, age, or anything that may be found in a clinical file.

Integrating sensor technology is crucial for developing effective personal digital twins. A personal Digital Twin may leverage several sensors, including medical devices, wearable sensors, and those contained in hand-held devices like smartphones (i.e. accelerometer, compass, light sensor, gyroscope, GPS, etc.) [4].

Chapter 4 leverages the design ecosystem to design and implement a proof of concept for a PDT of the human heart as a part of the present thesis; this DT integrates ECG signals and leverages ML to detect ischemic heart diseases; this work was published in [17]. In this instance, sensor data is used to detect physiological problems. However, ECG data may also assess mental well-being and stress levels. In [18], a study was conducted to determine the stress a VR videogame may impose on the user by leveraging ML to predict the individual's stress level. Similarly, when analyzed in the spatial domain, EEG signals may be used to assess stress levels by leveraging Convolutional Neural Networks (CNN) [19].

Up to this point in the present thesis, the terms Digital Twin (DT) and Personal Digital Twin (PDT) have been used interchangeably, as though they are synonymous. However, it's essential to recognize that while a PDT is indeed a specific case of DT, it comes with unique challenges. These complexities stem from the intricacies of digitally replicating a complex entity like a human being, even when the PDT focuses on copying a specific aspect of the person, such as the human heart. The following section will explore these challenges associated with PDTs.

2.2.1. Personal Digital Twin

Expanding the application domain of DT to human PTs (i.e. persons) is an idea introduced in [5] by El Saddik; this particular case of DT is known as Personal Digital Twin (PDT), and its design presents challenges specific to it; a PDT ought to have the following:

- An identifier to match each DT with a unique PT,
- sensors and actuators to capture data from the PT as well as to interact with the world,
- artificial intelligence (AI) so the DT can make sense of the data coming from the sensors,
- communication technologies (5G, Bluetooth, Wi-Fi and others),
- a way of representation (Multimodal Interaction) to act as an interface,

- trustworthy (to delegate important tasks) and
- private and secure.

It has been proposed that a DT would also be concerned with cloning the relations of the PT to other PT and objects in the PT environment in addition to the PT and its characteristics [20]. However, El Saddik raises an essential question regarding the design of PDTs: How can a PDT interact directly with the world or even with its own PT? This is a crucial aspect of this new paradigm; now, the PDT may communicate to the real world or even provide feedback to the PT directly, effectively influencing the world. Providing input to the PT is also a key feature of PDTs because it can affect the PT, enabling PDTs to be deployed for health and well-being [5]. An ecosystem called Dtwins, designed to create a digital copy of a person, which lays out all the necessary tools and technologies to implement a DT of a person, was introduced in [4].

A PDT is designed to clone and render different aspects of an individual's mental and physical states. Although this technology is still in its early stages, examples exist in the scientific literature.

The bio-medical PDT presented in [21] gathers information from multiple sources, including transcriptomic data (RNA sequencing), cellular data (related to metabolism, protein synthesis, replication and motility), organ information (related to organ function as part of systems) and exposomic data (connected to an individual's lifestyle, diet, exposure to toxicants, and more). The biomedical PDT is designed to be a general-purpose model for health as it provides an overview of various aspects of the human body and its ailments. The biomedical PDT has two main features: it detects diseases by leveraging Graph Neural Networks and can forecast results in clinical scenarios by leveraging GAN Networks. The authors of [3] discuss some characteristics of a "general-purpose" PDT for health. The hypothetical PDT for health is designed to create virtual copies of various organs and aspects of a person's health or even replicate single elements of human health, like body functions, organs, or even specific clinical conditions, such as diseases or other aspects of an individual's health. Additionally, the hypothetical PDT for health is expected to have aggregates of DTs to address health issues on a population scale. In [22], a use case for PDT in the healthcare industry is presented, where the PDT is used for trauma. In case of an accident, the PDT collects all data from the patient and relays it to the trauma physicians in the ER at the hospital. The premise is that the treating physicians have an action plan ready to help the patient when the patient arrives at the hospital. Another PDT for health, HospiT'Win, is designed to assist in hospital settings [23]. It supports clinical decisions by forecasting outcomes.

The previous examples have in common that they are designed for a particular use case. Also, data collection in these cases requires specialized medical devices, which are not readily available for domestic use. Hence, these PDTs lack the capability to extend beyond their particular use case.

Before delving deeper into PDTs for well-being, it is essential to understand the dynamics of health, well-being, and happiness. A PDT for well-being may be a more holistic approach to enhancing people's health and treatment outcomes.

2.3. Innovative Personal Digital Twins and their role in healthcare outcomes

Research works that fit the definition of Personal Digital Twin can be found in literature, even when researchers do not present their work as such. In the context of PDTs for health and well-being, eight categories for grouping PDTs were identified and shown in Table 2-1. Technology use, application domain, and target user group were the attributes by which the categories were defined. The following sections detail these categories, the common characteristics of a PDT framed within each one of these categories, and their potential impact on well-being and healthcare outcomes.

Table 2-1. Categories and descriptions of personal digital twins as identified in the literature.

Category	Description	Referenced Papers
VR for Training and Education	VR-based simulations for education and skills enhancement	[24], [25], [26], [27] and [28]
Cognitive Assessment and Rehabilitation	Digital platforms for cognitive function assessment and improvement	[29], [30], [31], [32] and [33]
Health and Well-being Monitoring	Continuous health monitoring using sensors and digital twins	[34], [35], [36], [37], [38], [39] and [40]
Empathy and Understanding Through Immersion	Immersive experiences to foster empathy and support	[41]
Mental Health and Psychological Support	Digital interventions for mental health and stress relief	[42], [43], [44] and [45]
Remote Healthcare Delivery and Telemedicine	Digital twins and technologies for remote health assessments and care	[46], [47] and [48]
Interactive and Assistive Technologies for Disabilities	Enhancing accessibility and support for individuals with disabilities	[49]
Innovative Interfaces for Health Interaction	New interfaces for health applications, including rehabilitation and education	[50], [51], [52], [53] and [54]

Virtual reality (VR) for training and education. This category encompasses using VR to simulate real-life scenarios for educational purposes and skills training for healthcare professionals. VR provides an immersive environment where users can interact with virtual patients or scenarios, enhancing learning and practical skills without the risks associated with real-life practice. A PDT in this category would simulate the user's interactions within a virtual environment, providing personalized feedback and adapting scenarios based on the user's performance and learning curve. Some PDTs in this category can track skill development over time, suggest improvement areas, and customize the training experience. PDTs in this category improve intervention outcomes, patient care, and fewer medical errors by providing a realistic and risk-free training environment, thus enhancing the competency of healthcare professionals. PDTs that fit this category are reviewed in the following paragraphs.

In [24], an MR-based surgical trainer system for minimally invasive fetal surgery is introduced. In the study, the authors successfully validated the use of MR for training in fetal laser minimally invasive surgery; the resulting system was dubbed realistic and considered relevant for building skills by participating surgeons during the study. In this case, the PDT is cloning the trainee's capacity, skills, and movements, rendering them inside an MR platform.

The authors of [25] developed and evaluated PoLaRS, a portable virtual reality training system for robotic surgery. The experimental evaluation involves exercises where participants were asked to sort marbles by colour and pick them up to feed them through circular cut-outs, emphasizing depth perception. These exercises were designed to test bimanual performance and the effectiveness of the PoLaRS system. On the other hand, the control group participants were exposed to a video on the da Vinci Xi system, another robotic surgery system [26]. The post-test was conducted on a da Vinci Skills Simulator unit connected to the da Vinci Xi console. An interesting observation from the study was the varied responses from participants when asked about the effectiveness of the PoLaRS system in improving their performance on the da Vinci system. Some felt that the PoLaRS system contributed to their performance, while others did not. Integrating such VR systems into the concept of PDTs can revolutionize medical training, allowing for more immersive, interactive, and realistic training scenarios. The PoLaRS prototype is presented as a potential training tool for robotic surgery, and the study aims to validate its effectiveness by comparing it with an established system, i.e., the da Vinci system.

Authors in [27] explore using immersive virtual reality (VR) for onboarding health care providers in a new clinical space. The primary objective was to familiarize staff with the latest clinical environment before transitioning. This was achieved by creating a "digital twin" of the new clinical spaces, which was experienced through immersive VR. The VR environment was designed to be realistic, interactive, and responsive, including virtual objects and avatars. This allowed participants to gain exposure to crucial unit locations, new room functionalities, and even practice skills as part of a clinical case. The work presented in this paper can be seen as a precursor to more expansive applications in virtual environments, where entire virtual hospitals or healthcare facilities might be created for various purposes.

A VR platform to train dental students is presented in [28]. The study aims to assess the educational benefits of using VR technology to teach dental students, focusing on home dental care practices. The authors designed VR-based teaching materials for a dentistry program. Students who participated in the VR-based training showed significant improvements in knowledge and confidence regarding home dental care practices. The study utilized hierarchical cluster analysis to demonstrate the change in students' understanding before and after exposure to VR teaching materials.

Cognitive assessment and rehabilitation. The PDTs in this category are centred around assessing and rehabilitating cognitive functions. It often integrates immersive technologies like VR or AR to provide engaging experiences that can measure or improve cognitive abilities. A PDT for cognitive assessment and rehabilitation would digitally represent an individual's cognitive performance, identifying deficits and monitoring progress over time. It could adapt rehabilitation tasks based on the user's performance, offering a personalized rehabilitation plan. Personalized cognitive rehabilitation can lead to more effective management of cognitive deficits, improving the quality of life for individuals with cognitive impairments. Enhanced cognitive function can also contribute to better daily functioning and independence, positively impacting overall well-being. PDTs that were designed for cognitive assessment and rehabilitation are discussed next.

In [29], a prototype for a cognitive assessment and rehabilitation platform (bWell) is presented; in this platform, the user may select from various VR environments with different tasks that may operate in assessment and rehabilitation modes. This platform aims to aid users in the management of deficits in cognitive function. The platform is designed for individual users (non-cooperative). In its architecture, authors do not report the collection of physiological signals during the interventions. However, the possibility is presented as a future improvement to the platform.

In [30], a platform with sensory biofeedback is developed, aiming to create a more immersive and effective therapy experience for alcohol addiction. The research addresses the increasing demand for psychotherapy, particularly for alcohol addiction, and the challenges faced due to long waiting times for treatment. The paper emphasizes the potential of VR-based exposition scenarios, which can create highly individualized or common trigger scenarios for patients. These scenarios are not reliant on the patient's imagination or real-life testing but can simulate environments that provoke cravings. The study involved conducting interviews with professionals in psychotherapy and treatment of addictions to ensure the practicality and relevance of the approach. The experimental study was conducted with ten healthy test subjects. The study considered aspects like usability or degree of immersion rather than evaluating the approach's effectiveness; it was a pilot study. Based on interviews with experts, a shopping scenario was chosen as it's a common critical situation for many alcohol-addicted patients. According to the authors, the cravings for addictive substances like alcohol cause similar reactions to the ones observed during episodes of acute stress, showing significant increases in heart rate and electrodermal activity, showcasing the value of integrating sensors capable of collecting such signals. The authors identified that this approach to managing addictions presents a high potential. However, the platform does not meet clinical requirements for real scenarios.

In [31], a system that integrates EEG signal processing with assistive technologies, such as an automated audio announcement system and virtual reality (VR), to aid individuals with disabilities, specifically visually impaired wheelchair users. Additionally, the study presents a secondary use case for the system on the health monitoring of soldiers using an FPGA-based system that processes EEG signals to detect abnormalities.

In [32], the authors explore the effects of a task-driven rehabilitation system on cognitive function, sensorimotor abilities, and brain connectivity in older adults. The study combines subjective scales, objective measures (such as EEG data), and behavioural indicators to evaluate the platform. The findings suggest that VR-based rehabilitation can improve cognitive function and sensorimotor abilities in elderly individuals, potentially offering a more effective and personalized approach than traditional rehabilitation methods.

A stroke assessment platform for patients during stroke rehabilitation is introduced in [33]. Stroke patients often experience motor impairment, and rehabilitation is crucial for recovery. Customizing rehabilitation tasks is essential for enhancing the effectiveness of these interventions. Rehabilitation robots are being developed to alleviate the workload on physiotherapists and offer accurate assessments of patient recovery. However, patient evaluation and robotic rehabilitation task design rely heavily on physiotherapists. This digital twin captures the patient's motor control process, including planning, neural impairment, and muscle activation. By adjusting the state and parameters of this model, a digital representation of the patient is created and updated, reflecting their adaptation and progress in rehabilitation. This digital twin is then used to plan customized rehabilitation tasks. The research indicates significant potential for using digital twins in stroke rehabilitation to create personalized, effective rehabilitation tasks. Future work could include integrating multimodal data and exploring reinforcement learning for task planning.

Health and well-being monitoring. This category involves continuously tracking and monitoring health parameters and behaviours using sensors to transmit data to a PDT, aiming to provide insights into an individual's health status. The premise is that a PDT would collect and analyze data from various sources, including wearable sensors and environmental inputs, to create a comprehensive health profile. It would identify patterns, predict health outcomes, and recommend maintaining or improving health. Continuous monitoring can lead to early detection of potential health issues, enabling timely interventions. This proactive approach to health management can prevent complications, reduce the need for emergency care, and improve overall well-being.

The research presented in [34] outlines an initiative focused on leveraging the Digital Twins (DT) concept to enhance the monitoring and management of chronic diseases. The central premise is that continuous and effective monitoring of chronic conditions and their treatments can significantly reduce risks and improve patients' quality of life. However, achieving this level of monitoring requires innovative engineering systems capable of supporting these advanced capabilities. This work proposes using Digital Twins, virtual representations that mimic the structural elements and dynamics of any physical asset (in this case, a patient) throughout its lifecycle. The application of DTs in healthcare is seen as a significant advancement toward improving and tightening the interactions between systems, caregivers, and patients. By creating a digital counterpart of a patient, healthcare providers can gain insights into the patient's health status, treatment responses, and potential outcomes in real time, facilitating a more personalized and precise approach to medicine. Furthermore, integrating data-driven methods, such as Machine Learning, with Digital Twins is highlighted as a critical mechanism for continuously tracking patients' health and evaluating medical treatments virtually. This approach allows for the simulation of treatment outcomes, adjustment of treatment plans based on real-time data, and the anticipation of potential health risks before they manifest severely.

In [35], a home-based monitoring CPS that uses fuzzy rules to calculate a “risk code” corresponding to the patient's current state is presented. The patient's health status can be consulted by treating physicians, family members, etc. The patient data is stored temporarily on the patient's mobile device (smartphone) and permanently in a cloud-based repository accessible through a web service (WS).

A prototype for a PDT for remote health monitoring is presented in [36]. The study focuses on the role of DTs in healthcare and proposes a generalized DTs architecture. This architecture monitors various aspects such as social distancing, queue status, table occupancy, and worker tracking during the pandemic. The architecture is layered, including physical sensors, data aggregation, connectivity for information exchange, and the DTs layer, which includes the virtual representation of the physical world. The implemented prototype also consists of a VR interface showing the obstacles for the DT navigating its environment.

A novel framework for elderly healthcare management using digital twin technology was introduced in [37], dubbed CloudDTH (Cloud Digital Twin Healthcare). This framework leverages advancements in big data, cloud computing, and the Internet of Things (IoT) to bridge the gap between the physical and virtual domains in healthcare. This study focuses on creating a digital twin to monitor, diagnose, and predict individual health trends, particularly the elderly, through wearable medical devices and other digital tools. The proposed CloudDTH framework addresses two primary challenges in the era of precision medicine: the first is achieving personal health management throughout the entire lifecycle of elderly patients, and the second is the dynamics of the physical and virtual worlds to achieve the ideal of smart healthcare. To tackle these challenges, the paper outlines a comprehensive approach involving integrating digital twin technology within a cloud-based healthcare system. This integration facilitates a seamless interaction and convergence between the physical and virtual medical care spaces, enhancing the precision and efficiency of healthcare services.

Authors of [38] introduced MedicaLog, a health data platform. Personal Digital Twins have the potential to play a central role in the modern healthcare ecosystem, offering a centralized solution for managing the vast amounts of health data; data protection and privacy must be at the core of PDTs for health and well-being. These platforms enable a significant shift towards preventive healthcare, allowing for the early detection of potential health issues and the development of personalized care plans. By engaging patients directly and providing a secure environment for their health data, such platforms empower individuals to take an active role in their health management. This approach streamlines the complexity associated with health data interpretation and allows for a more personalized, proactive, and data-driven delivery of healthcare services.

A Personal Digital Twin for Parkinson’s early detection is presented in [39]. The research introduces a Digital Twin-Based Healthcare System, focusing on improving healthcare for elderly and disabled individuals, particularly in remote areas. Some health conditions present a progression over time, especially chronic health conditions such as Parkinson’s disease, which show a constant progression and would benefit from continuous monitoring and assessment to gauge its progression accurately. The platform is designed to offer earlier prediction and patient-centric diagnosis of diseases like Parkinson's through constant monitoring and smart virtual care facilities. The Optimized fuzzy-based k-nearest Neighbor (OF-k-NN) classifier model is central to the platform design, which uses voice data to predict Parkinson's disease with notable accuracy and efficiency. By leveraging digital twins and IoT in healthcare, the platform offers a promising approach to providing personalized, efficient, remote monitoring and diagnosis services.

An IoT-based PDT to enhance healthcare procedures for diabetic retinopathy (DR) is introduced in [40]. The authors propose a framework incorporating these technologies to develop a digital replica of a patient's health status, enabling more efficient and precise healthcare services. As described in this paper, integrating Digital Twin and IoT in healthcare represents a substantial advancement in eHealth. This research contributes to the growing knowledge of applying AI and machine learning in medical diagnosis and continuous patient monitoring. It also aligns with the trend towards more personalized medicine, where treatment plans are tailored to the individual rather than adopting a one-size-fits-all approach. The results indicate that the model could be used in real-world scenarios to support ophthalmologists and other healthcare providers diagnose DR, potentially leading to earlier interventions and better patient outcomes. The use of transfer learning with EfficientNet in the study highlights the potential for utilizing pre-trained models to accelerate and enhance the development of healthcare applications. Overall, this paper shows the feasibility of creating a digital twin for healthcare purposes, opening possibilities for its application in other medicine and health monitoring areas.

Empathy and understanding through immersion. This category focuses on producing immersive technologies to create experiences that foster empathy and understanding towards specific conditions or situations, aiming to support and improve societal attitudes. A PDT in this category would capture the PT's perspective to render it for others to see or even experience it from the PT's point of view, leveraging the interrogative nature of the PDT. A way for others to "walk in the shoes" of the PT, providing insights into their challenges and perspectives. It would adjust scenarios based on user interactions to deepen understanding and empathy. By fostering empathy, such PDTs can contribute to a more supportive and inclusive society. For individuals experiencing the point of view of the PT, this increased understanding can lead to better social support, reducing stress and improving mental health outcomes. Some examples in this category are presented next.

A PDT presenting the PoV of a breastfeeding mother is introduced in [41]; the platform was dubbed “Virtual Feed,” a virtual reality (VR) simulation of the experience of breastfeeding in public spaces. The research seeks to understand how the platform can foster empathy and understanding towards breastfeeding mothers, especially in public spaces. The researchers employed a participatory design approach, involving breastfeeding mothers in the design process. This ensured that the simulation was grounded in real-life experiences. "Virtual Feed" places users in the perspective of a breastfeeding mother in a public setting, such as a café. The simulation presents various scenarios and challenges a mother might face, such as public scrutiny, finding a comfortable breastfeeding place, and managing the baby's needs. The simulation is designed to be immersive, allowing users to engage fully with the mother's perspective. The authors concluded that the platform fosters understanding and empathy, potentially leading to better support systems for breastfeeding mothers.

Mental health and psychological support. A PDT in this category is focused on offering therapeutic interventions, stress relief, and coping strategies for mental health issues. Sometimes, a PDT in this category may use VR to render immersive therapeutic environments. A PDT for mental health would offer personalized therapeutic interventions based on the user's emotional state and mental health history. PDTs in this category may track progress, adapting interventions to the user's changing needs. Personalized mental health support can lead to more effective management of mental health conditions, reducing symptoms and improving quality of life. This tailored approach can enhance the accessibility and effectiveness of psychological support. Instances of this category are analyzed next.

The platform "Trash It, Punch It, Burn It" is a VR platform to help individuals cope with negative thoughts, particularly those triggered by texts such as emails, SMS, and WhatsApp messages, are presented [42]. Many people frequently experience negative thoughts, which are linked to severe mental health problems. While there are therapeutic interventions to manage these thoughts, they often require clinical expertise, making them less accessible for everyday use. The study explores how technology, specifically virtual reality, can offer new possibilities to support coping strategies for negative thoughts. The authors developed a VR prototype as a smartphone application combined with a cardboard VR headset. The minimalistic virtual environment focuses on the text to help users concentrate on the message rather than environmental elements. The prototype was designed for brief everyday interventions, making it easily accessible without technical barriers. The study collected qualitative data, and several themes emerged, including shifts in thoughts and emotions regarding the messages, feedback about the prototypes, gestures and embodied cognition, VR, and integration into everyday life. Most participants believed the VR intervention positively affected their perception of the negative message, making them more relaxed and reflective.

Authors in [43] present a platform that enables users to decorate their own Virtual Environment (VE) using various brushes, environments, and manipulable objects. This application was found to be beneficial for emotional expression. The following steps involve exploring guiding VR using qualitative visualizations and voice-based guidance. In this work, the authors examine the potential of VR in fostering mental well-being, which is a crucial aspect of health. The findings and methodologies in this research can be applied to the broader context of creating a personal digital twin to monitor, guide, and enhance an individual's mental health.

A prototype for a procedurally created VR maze with exercise rooms is presented in [44]. The mazes are procedurally generated by a neural network trained with a Deep Reinforcement Learning (DRL) algorithm. The goal is to adapt the difficulty of the mazes to the player's capabilities, balancing physical exertion and cognitive challenge. This platform effectively creates a PDT by collecting data directly from the participant and adapting the challenges accordingly. The authors collected data on participants' experiences before and after the experiment, including the Game Experience Questionnaire (GEQ). Heart rate was also collected to gauge exertion levels; in the study, the authors proved the viability of having exergames as part of VEs successfully integrated exergames in a Virtual Environment of exploratory nature; in turn, it has been proven that exergames can have a positive impact on mental health.

A platform for emotional state assessment was introduced in [45]. The integration of emotion recognition systems with PDTs is a significant advancement. It broadens health monitoring to include psychological well-being, enhancing patient engagement and improving the efficacy of remote care by providing critical emotional insights. Leveraging machine learning (ML) and deep learning (DL) technologies, an emotion recognition system may capture and analyze facial expressions in real-time, facilitating the creation of personal digital twins that reflect patients' emotional states, which presents the possibility for mental health professionals to tailor psychological support and interventions to the unique needs of each patient. Such a system would be able to detect a broad spectrum of emotions under varied conditions accurately. A PDT enabled with non-invasive, real-time emotion recognition enhances the understanding of patient emotions, leading to the early diagnosis of psychological conditions, more accurate assessments of patient well-being, and the implementation of personalized interventions.

Remote healthcare delivery and telemedicine. PDTs in this category enable remote health assessments, consultations, and interventions, leveraging digital twins for healthcare delivery across distances. This category addresses the need for accessible healthcare services regardless of geographical constraints. A PDT in this domain may facilitate remote monitoring and consultations, integrating health data from multiple sources to provide healthcare professionals with a comprehensive view of the patient's health. It could also support decision-making by predicting health outcomes based on current trends. Remote healthcare delivery can enhance access to healthcare services, particularly for underserved populations. It can reduce travel time and costs for patients, improve monitoring of chronic conditions, and enable timely interventions, leading to better overall healthcare outcomes.

The platform DiCRAs was introduced in [46] as an evaluation framework for the remote evaluation of AR-mediated interactions, particularly in health behaviours, the authors conducted two studies: a remote unmoderated between-users experiment and a remote moderated exploratory technology-probe study. The overarching goal of the research was to understand the expanding opportunities and challenges of remote studies in influencing personal health behaviours. The authors emphasize the potential of remote studies in shaping the future of extended reality (XR) interventions for emotional health while also addressing concerns related to privacy, selection bias, and data validity.

A platform for gait analysis for telemedicine is presented in [47]. The researchers developed and evaluated a virtual reality (VR) system for gait analysis, a significant aspect of telemedicine. The study explores the potential of VR technology to assist in diagnosing and treating gait-related health issues. It involves a prototype VR telemedicine system that allows for asynchronous gait analysis, meaning that the data can be captured and analyzed at different times and locations, which is particularly beneficial for remote healthcare delivery. The study presents a qualitative evaluation of the platform. The results indicated that many participants found medical terminology and the understanding of their pathology challenging. However, using 3D visualization tools was beneficial for understanding their pathology and underlying anatomy. The study showed that VR systems may provide a more comprehensive and accessible view of a patient's health, where immersive experiences can be utilized for health monitoring, diagnosis, and treatment planning.

Researchers in [48] presented a framework for an AI and IoT-based platform for oncology patient diagnosis. The research introduces an innovative Smart IoT Platform that uses augmented reality (AR) and artificial intelligence (AI) to enhance the diagnosis processes in cancer treatment, marking a stride towards achieving a Human Digital Twin. The proposed architecture integrates various components: a mobile application for data acquisition and patient interaction, a cloud database for data storage and retrieval, AR for visualization of patient data and MRI scans, and a Convolutional Neural Network (CNN) for image analysis. Significantly, the article discusses the preliminary steps in developing this framework and the potential impacts on healthcare delivery. The proposed system represents an evolution in patient care, potentially transforming oncology diagnosis and treatment through high-precision digital twins. Integrating AI with IoT provides a multifaceted platform that facilitates diagnosis and offers predictive insights into patient health, underscoring a shift toward more personalized and preemptive healthcare strategies.

Interactive and assistive technologies for disabilities. This category of PDTs focuses on enhancing accessibility and support for individuals with disabilities through the integration of digital twins with assistive technologies. A PDT in this category can provide customized support based on the individual's needs and abilities, integrating assistive devices and technologies to enhance daily functioning and independence. By addressing the unique challenges faced by individuals with disabilities, PDTs can significantly improve their quality of life, fostering greater autonomy and participation in society. Enhanced accessibility and support can lead to better health and well-being outcomes. Instances of PDTs in this category are analyzed next.

Researchers in [49] present a photoplethysmography (PPG) system, which utilizes sensors to detect and differentiate finger-level gestures. PPG sensors measure the volumetric variation of blood circulation, and the paper explores how these sensors can be leveraged to recognize gestures, especially those from the American Sign Language. The study found that it's possible to differentiate nine elementary finger-level gestures with an average recognition accuracy of over 88%. The paper acknowledges the challenges posed by factors such as skin colour and the impact of strenuous exercises on the PPG signals. The authors also express their intention to address these challenges in future work. A Digital Twin for well-being would benefit from gesture-based interactions, especially in physical rehabilitation; additionally, measuring blood volume with PPG-based techniques would enable indirect measurements of physiological signals related to the heart rate.

Innovative Interfaces for healthcare delivery. A PDT in this category explores novel interfaces (such as BCI, AR, and VR) for health-related applications, including rehabilitation, patient education, and interaction with healthcare systems. A PDT employing innovative interfaces can provide intuitive and accessible ways for individuals to interact with health information and services. It could adapt to the user's preferences and abilities, offering personalized experiences that enhance engagement and effectiveness. Innovative interfaces can make health-related information and services more accessible and engaging, leading to increased participation in health management. This can improve adherence to treatment plans, enhance patient education, and ultimately lead to better health outcomes. Instances of PDTs fitting this category are presented next.

Researchers [50] developed and evaluated a low-cost BCI interface to exert control over objects inside a virtual environment. The environment presents a simple challenge using left and right sensorimotor rhythms mu and beta to control the rotation of a single figure without moving the camera's point of view; when positioned correctly, the figure shows a Celtic symbol. The challenge is designed for a single user to solve it. In the study, the authors highlighted that most users could solve the challenge, thus offering an argument for using BCIs to interact in virtual environments. The study also showed the feasibility of collecting and using EEG signals in real-time applications for virtual environments.

NeuroDot is presented in [51] as a portable system with a custom VR headset for neuro-optical diagnostics; the research addresses the challenges and limitations of traditional EEG methods used in visual evoked potentials (VEPs) for clinical assessment. NeuroDotVR offers a solution to these challenges by providing an objective measure of neuro-ophthalmic performance without requiring active participation from the subject. This is especially beneficial for trauma patients who might find it challenging to offer subjective assessments. The system also explores the feasibility of using VR for Vision Therapy, allowing for monitoring therapeutic benefits and eye movement control. A personal digital twin for health and well-being would require real-time data and diagnostics to accurately represent and predict health outcomes. As presented in this paper, the advancements in portable neuro-optical diagnostics can contribute to developing comprehensive digital health profiles by bridging the gap between digital-based and real-world health diagnostics and interventions.

Therapy Tracker is presented in [52] as a cross-platform application for medication management. This application integrates augmented reality (AR) and machine learning (ML) technologies to enhance the user experience, particularly for seniors. The app manages a list of medications, dosages, refill information, and prescribing documents. It alerts elderly users about their medications by scanning their prescriptions using the smartphone’s camera. The app features a virtual assistant that guides users through visual and auditory processes, such as identifying the correct medication from a group of bottles. The paper also presents an architectural model, providing a foundation for other researchers to extend this work. In broader terms, implementing a Digital Twin for well-being involves creating digital representations of people’s health and the treatments and interventions to improve the individual’s health. "Therapy Tracker" aligns with this concept by using AR to virtually guide users in the real world and ML to process and predict medication-related tasks.

A prototype that leverages AR and smart glasses to assist physicians in treating chronic wounds is introduced in [53]. The platform is designed to document wound treatment procedures and characteristics of the wound. The application allows them to establish procedures hands-free while performing them. The study also investigated audio-based and physical interaction with the smart glasses.

The authors of [54] present a 360 Augmented Reality Visualization Platform (ARVP) as an adjunctive tool for patient education in neurosurgery. The researchers conducted a pilot study with neurosurgical patients who used the 360 ARVP and provided insights into the potential benefits of AR in patient education.

2.4. Analysis of existing research and the need for further exploration

This section identifies key areas where existing literature indicates a gap in research, highlighting opportunities for further study. It aims to outline the current limitations in our understanding and suggest directions for new research that can advance the field.

Comprehensive real-time processing design ecosystems. The literature reveals that there are currently no design ecosystems or frameworks that can be used as a starting point when designing a PDT for well-being or health. There is a need for a design ecosystem that acts as a guideline and informs the design of new PDTs for well-being, and that is capable of collecting and storing data locally while processing and generating intelligent feedback in real-time. In addressing this gap, the proposed ecosystem seeks to establish a novel benchmark in the seamless orchestration of PDTs for well-being, leveraging the capabilities of edge computing and ML.

Integration of diverse health data sources. The literature lacks detailed exploration into integrating heterogeneous health data sources within a single PDT framework, such as ECG signals, other physiological parameters, and self-reported symptoms. This thesis proposes innovative methods for the holistic integration of these diverse data sources, aiming to enhance the accuracy and comprehensiveness of well-being and health state inferences.

Adaptability and scalability of PDT ecosystems. The rapid evolution of technology and the dynamic nature of healthcare demands necessitate PDT ecosystems that are both adaptable and scalable. The current body of literature offers limited insights into how PDT ecosystems can evolve alongside technological advancements and scale to accommodate diverse healthcare scenarios. This thesis addresses this gap by proposing a flexible and scalable architecture for the PDT ecosystem, ensuring its relevance and efficacy in the face of technological progress.

Closed Feedback Loop Integration. A significant gap in the literature pertains to the integration of closed feedback loops within Personal Digital Twins (PDTs), particularly from an information theory perspective. Closed feedback loops are fundamental in dynamically adjusting systems based on the system output or the system's state, enhancing accuracy, efficiency, and user quality of experience. In the context of a PDT ecosystem for health monitoring, the literature lacks a comprehensive exploration of how closed feedback loops can be effectively implemented to refine health predictions, adapt interventions, and personalize health monitoring in real-time. This thesis aims to bridge this gap by leveraging principles from information theory to design and implement a closed feedback loop mechanism within the PDT ecosystem. Such a mechanism would allow for the continuous adjustment of health monitoring and intervention strategies based on real-time data analysis and user feedback, thus ensuring the system's adaptability and responsiveness to changing health conditions.

As mentioned in this document, in the present research, a set of guidelines and design principles for designing PDTs for well-being have been identified; these are bundled into a design ecosystem dubbed TMW. Chapter 3 introduces the proposed ecosystem, which aims to bridge the research gaps identified in the preceding discussion. Chapter 3 lays the foundation for a comprehensive understanding of how the ecosystem integrates real-time processing, edge computing, and closed feedback loops, among other innovative solutions, to advance the research of Personal Digital Twins for well-being.

Chapter 3

3. Towards a design ecosystem for a personal digital twin for well-being

---

The design of personal digital twins for well-being presents several challenges outlined in the previous chapter. This chapter introduces a design ecosystem for PDTs for well-being. This ecosystem solves said challenges by introducing general design guidelines and base structures and modules that facilitate the design and implementation of new PDTs for well-being using the ecosystem as a starting point.

One key aspect of this design ecosystem is integrating a closed feedback loop mechanism for self-regulation and adaptation. In information theory, closed feedback loops are channels through which system outputs are routed back as inputs. This process of recursive information flow allows for correction and optimization, enhancing the system's overall function. The present chapter argues that a PDT for well-being would require a closed feedback loop; such a feature would improve the platform's effectiveness in enhancing the PT's well-being.

3.1. Ecosystem for Personal DT

Personal Digital Twins (PDTs) are at the convergence of AI, Multimedia, and XR. An essential aspect of PDTs is their ability to provide the physical twin (i.e., the person) with valuable insights and recommendations according to the purpose of the PDT. For example, a health-oriented PDT would give the PT helpful feedback on its health state, while a sports-oriented PDT would provide feedback on the person’s physical aptitude.

The design of Personal Digital Twins (PDTs) is a multifaceted challenge that necessitates a comprehensive approach. A reference ecosystem offers a structured framework that benefits the research community.

A design ecosystem sets standardized guidelines and protocols to ensure interoperability and compatibility across PDT systems. This standardization facilitates a unified approach to development. It also consolidates established best practices, enabling researchers to build upon proven methodologies. This helps deliver robust and reliable PDT solutions while avoiding common development challenges. A shared reference ecosystem encourages collaborative efforts among researchers. It serves as a common ground for exchanging ideas and findings, which can lead to more efficient and innovative outcomes.

A design ecosystem also provides a template for replicable research, vital for scientific validation. It offers a scalable model that can incorporate advancements and new data sources. In turn, the ecosystem allows for establishing benchmarks to evaluate PDT systems. Researchers can gauge the performance of their designs against these benchmarks to ensure continual improvement. A design ecosystem can streamline the development process by providing foundational elements, allowing researchers to focus on innovation rather than initial setup complexities.

The design of PDTs is inherently multidisciplinary. An ecosystem facilitates the integration of diverse fields, ensuring that designs benefit from a comprehensive array of expertise.

Introducing a design ecosystem such as TMW has additional benefits. By embedding guidelines for regulatory compliance and ethical considerations into the ecosystem, PDTs adhere to necessary privacy, data protection, and user consent standards.

By leveraging a reference ecosystem, researchers can effectively navigate the complexities of Personal Digital Twins design, leading to more sophisticated, user-centric, and ethically responsible solutions.

Part of the effectiveness of Personal Digital Twins (PDTs) in providing tailored insights hinges on the Quality of Experience (QoE), which profoundly influences the physical twin's perception of the PDT. A positive QoE not only bolsters the adherence and adoption of the PDT by enhancing user satisfaction but also amplifies the impact of the PDT's recommendations, leading to more meaningful improvements in the physical twin's well-being.

Quality of Experience (QoE) is a user-centric measure of overall satisfaction with a service or product. In the context of the design ecosystem for PDTs, a QoE-driven feedback loop is used to refine the AI-inference engine and multimodal interactions. The QoE feedback ensures the PDT remains aligned with user preferences and expectations. The system can adapt to changes in user behaviour or circumstances.

One could argue that a Quality of Experience-driven closed feedback loop is necessary for the success of Personal Digital Twins. Through the continuous incorporation of user feedback, PDTs can offer a truly personalized and dynamic experience, fully realizing the potential of PDTs in enhancing individual lives [55], [56].

Integrating this QoE closed feedback loop into the design of a PDT ecosystem brings multifaceted benefits. It champions a user-first approach, aligning with the core objectives of individualized health and well-being enhancements. By evolving through user interactions, the PDT system becomes an organic entity, fine-tuning itself in line with user-driven data. This responsive feedback is invaluable, allowing the system to anticipate and adapt to user needs proactively. It becomes instrumental in making informed decisions for feature enhancements and system upgrades. Moreover, it provides benchmarks for performance against user expectations, maintaining the system's relevance and competitive edge.

An active QoE loop can significantly reduce user drop-off rates by promptly addressing concerns and improving the overall experience. Such a loop, intrinsic to the PDT ecosystem, is indispensable for nurturing long-term user engagement, optimizing system functionality, and bolstering trust in the technology. The Quality of Experience (QoE) closed feedback loop is critical in personalizing interventions through PDTs for well-being.

This thesis also identifies a set of components, which are necessary to enable a PDT for well-being to deliver tailored interventions. The design ecosystem has three significant structures dubbed modules: Data Source, AI Inference Engine, and Multimodal Interaction.

The three modules of the design ecosystem are detailed next; each one of these modules groups a set of components, the relationships they maintain among these components, and how the data flows between them. Figure 3-1 is an overview of the PDT design ecosystem for well-being.

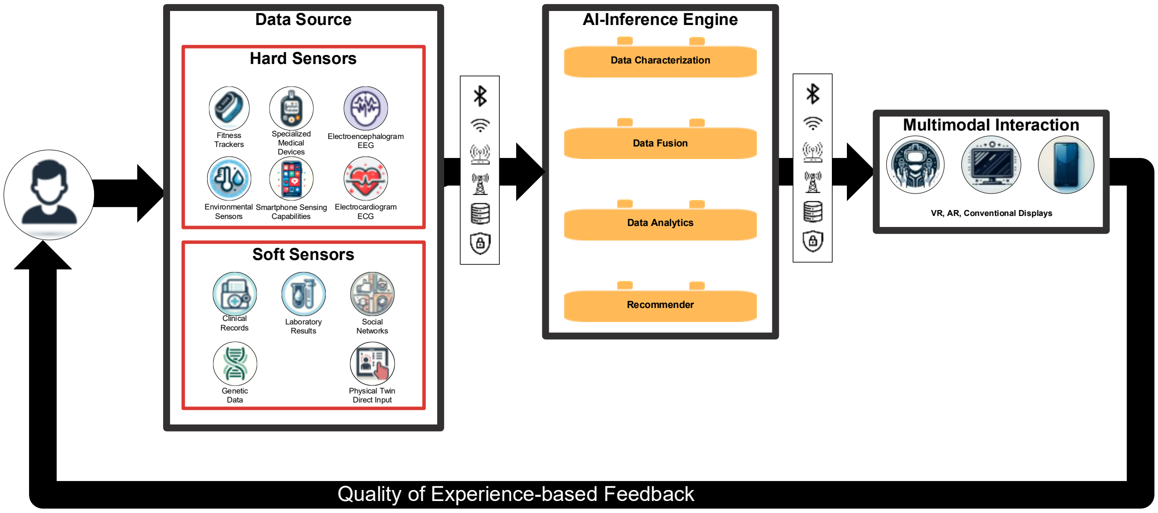


Figure 3-1. Diagram of the proposed design ecosystem for personal digital twins for well-being.

The design ecosystem follows a pipeline configuration similar to most modern data solutions, except that the ecosystem is paired with a closed feedback loop that presents the benefits already mentioned in this chapter.

Data Source. This module collects data from the PT, sensors on the PT, environmental sensors, if available, clinical records, and social networks. Sensors may be grouped into two categories: hard and soft.

- **Hard sensors**. These sensors are essential tools that meticulously monitor physiological and environmental parameters. Devices like fitness trackers, worn on the wrist, continuously gather data on cardiovascular activity, physical exertion, and sleep patterns, providing insights into our daily health. Environmental sensors further safeguard well-being by monitoring air quality and temperature, alerting us to potential hazards like pollutants and extreme conditions. Specialized medical devices, such as glucose monitors, electroencephalograms (EEG), and electrocardiograms (ECG), offer deeper health insights by tracking brain and heart activity. These advanced sensors detect early signs of neurological and cardiac conditions, facilitating timely medical interventions. Smartphones come with various built-in sensors that enable them to track movement, orientation, and location through accelerometers, gyroscopes, and GPS. They can also measure environmental attributes like light and atmospheric pressure with ambient light sensors and barometers. Some models include health monitoring capabilities, such as heart rate sensors. Additionally, microphones and cameras provide audio and visual inputs that support everything from augmented reality to environmental analysis. Through this interconnected system of hard sensors, our health is monitored and intricately understood, helping us maintain optimal well-being.

- **Soft sensors**. In contrast to hard sensors, these types of sensors capture data from non-physical sources, such as digital documents and online interactions. They primarily leverage the information available through digital footprints like medical records and social media activity. For instance, by analyzing an individual's interactions and behaviours on social media platforms, soft sensors can infer mental health states and social well-being. This kind of data, rich in qualitative insights, provides a deep understanding of a person's psychological environment and social dynamics. Additionally, medical records serve as critical soft sensors by providing a historical context of an individual's health, including past medical diagnoses, treatment outcomes, and genetic information. In addition, the physical twin may also input data directly, such as self-reported symptoms or answers to on-screen surveys; this data can offer an insightful view of the PT’s self-perceived physical and mental state. Their integration with hard sensors provides a holistic view of an individual’s health and behaviour.

AI-Inference Engine. This module pre-processes and feeds the data to the Data Analytics component (AI). The Data Analytics component analyzes the pre-processed data and extracts valuable information. Finally, the Recommender component enables the PDT to make recommendations that can improve the health and well-being of the person. All components inside the AI-Inference function like a data pipeline would:

- **Data Characterization**. Data collected in the data source must be summarized and processed to extract meaningful data features (i.e. characteristics) about the PT's well-being or behaviour. This component operates individually on each data stream from the data source module.
- **Data fusion**. Following the pipeline, this component is responsible for arranging the features extracted from different data sources and preparing the data to be fed into the analytics model. Depending on how many models, their inputs and the format of the inputs, this component would combine the data features having a corresponding number of outputs.
- **Data analytics**. This component initiates AI processes for data analysis using machine learning and deep learning techniques to identify patterns and make predictions. It processes biological, clinical, environmental, and social networking data to forecast well-being issues and assess social interactions. The module also analyzes direct measurements from fitness trackers, ECG, and EEG sensors to predict psychological states and infer current health conditions in real-time.
- **Recommender**. This module leverages data analytics results to generate personalized recommendations for the PT, aiming to improve its well-being. Continuously monitoring the real twin's health, it uses real-time data and AI techniques to offer suggestions that promote well-being. However, the suggestions must consider the PT circumstances to encourage adherence and adoption, thus improving healthier behaviours and well-being outcomes.

Multimodal Interaction. This module is responsible for digitally representing the Physical Twin (PT) by displaying the Personal Digital Twin (PDT) alongside the PT's health status and the inferences and recommendations generated by the previous module, the "AI-Inference Engine." This is how the PDT interacts with the physical world and its PT, creating a feedback loop driven by the Quality of Experience (QoE) philosophy.

Privacy and Data Protection. In the design ecosystem of Personal Digital Twins (PDTs), robust security measures are paramount, ensuring that data access is strictly controlled and limited according to the physical twin authorization. When data needs to be transmitted over the internet or a public network or recovered from a given end-point, a practical approach is to implement token-based authentication; using JSON Web Tokens in modern REST APIs is a good example[3]. This ensures that data transactions between the system's components are securely authenticated and that only intended entities (i.e. modules of the PDT) can access data when said data must be transmitted over the internet or public networks. Traditional password methods, which do not necessarily verify user authenticity, are thus replaced by more reliable token-based mechanisms. In addition to token-based authentication, the PDT system employs advanced encryption techniques to safeguard the confidentiality of data during its transmission across the network. This is essential to prevent unauthorized access and data breaches. Privacy settings within the PDT are highly customizable and intelligently designed to accommodate the varying sensitivity of the information handled. The system can discern which data types to share based on predefined privacy profiles, which can be adjusted depending on the scenario—from sharing employment details with professional contacts to requiring active consent for transmitting health data to medical services.

Communication. Communication within the PDT ecosystem is designed to optimize the Quality of Experience (QoE). Different technologies are employed to ensure efficient and secure communication depending on the type and volume of data being exchanged. For instance, high-volume data transfers, such as those needed for haptic-mediated interactions, might utilize high-speed internet like optic fiber, 5G, or others offering similar speeds, whereas simpler data

---

[3] JWT, or JSON Web Token, is a compact, URL-safe means of representing claims to be transferred between two parties. It consists of three parts: a header, a payload, and a signature, each separated by a dot ('.'). The header typically contains the token type and the signing algorithm being used, while the payload holds the actual claims and additional data such as the issuer, expiration time, and subject. In JSON APIs, JWTs are used for authentication by encoding user credentials and permissions, which the server verifies for each request to ensure the user has the correct access permissions, typically by checking the token's signature to confirm its validity.

exchanges, like audio communications, might only require 4G. IoT devices within the system typically use Bluetooth for local data transmission.

This thesis explains health as an integral state of well-being encompassing physical, mental, and social well-being and recognizes the significant impact of lifestyle and environmental factors on it. These three forms of well-being are interrelated; when one is affected, the other two may also be influenced by a domino effect. A PDT for well-being would contribute to a favourable outcome by detecting a condition in the early phases. Going even further, a PDT for well-being could also recommend changes to lifestyle or treatments for the PT to undergo, which would consequently improve its health.

Some factors and parameters help either predict a risk of a future health problem or to know a person's current state of health. These are mentioned in [2] and may be categorized into seven groups: physical health, lifestyle, mental and psychological health, socioeconomic, gender, context, and culture. While some of these factors may impact one of the three components of health, others will affect more than one. At the same time, some of these parameters may indicate a current health problem and, at the same time, be correlated with an increased risk of future health problems. Take, for example, the vitals, more specifically, blood pressure. High blood pressure may indicate heart disease, a current health issue. At the same time, uncontrolled high blood pressure can lead to future health problems like kidney damage. Going even further, the combination of a good diet, sleeping habits and exercise can help prevent (non-ischemic) hypertension in some cases. In this case, lifestyle factors have an impact on physical well-being and mental well-being, too. In other words, the PDT for well-being must be capable of collecting data on dietary habits, like calorie intake. In [57], [58] machine learning techniques are applied to estimate calorie intake. Human Activity Recognition (HAR) is a prolific field of research that deals with in which that has given us several research and academic products that report different techniques for recognizing human activities such as sleep, running, walking, sitting, and resting; the research in [59], [60] are great examples.

The PDT for well-being can use, improve, or extend any existing techniques in HAR to track calorie input and output as well as sleep patterns. This would allow the DT for Health to make recommendations based on the collected data and promote health for its PT.

The following sections detail the behavioural aspects of the proposed design ecosystem (i.e. TMW) and provide a component diagram that will enhance the discussion by presenting the structural elements of TMW. These diagrams will help to understand the mechanisms through which PDTs for well-being analyze, process, and utilize data to improve the Physical Twin (PT) health. These diagrams meticulously map out the behaviour of PDTs and the data flow within the system, offering a visual representation of how data is captured, analyzed, and transformed into actionable health recommendations. This step-by-step breakdown highlights the design ecosystem’s operational logic.

The first diagram is the Level 0 Data Flow Diagram, which will provide an intuitive view of how the data will be collected from the PT into the PDT for well-being and back to complete the closed feedback loop.

3.2. Level 0 data flow diagram

A Level 0 Data Flow Diagram (DFD), also known as a Context Diagram, represents the simplest form of DFDs used in system analysis and design. This diagram serves as an entry point into understanding the overall system, providing a high-level overview of a system's main processes and the flow of information between them and external entities. Here are the key elements of a Level 0 DFD:

- Single Process Node: Unlike more detailed DFDs, the Level 0 diagram typically contains only one major process, which is depicted as a circle or a rounded square. This process node represents the entire system as a single process without going into the internal details.

- External Entities: These are entities outside the system that either send data to or receive data from the system. External entities can be users, other systems, or external organizations, represented by rectangles or squares on the diagram.
- Data Flows: Arrows connecting the entities and the system illustrate the flow of data into and out of the system. These data flows show what type of data is being exchanged between the system and its external entities.

A Level 0 Data Flow Diagram (L0DFD) helps understand a system's broad interactions and data flow. It provides a high-level system overview, delineating how data flows between external entities and the system. This diagram helps define the system's scope and boundaries, showcasing the significant processes involved and their interaction with external factors such as users (the PT in the present case), other systems, or data services.

The utility of a Level 0 DFD lies in its simplicity and clarity. It allows stakeholders, regardless of their technical background, to grasp the fundamental operations of the system without getting overwhelmed by detailed functionalities. This introductory diagram is especially valuable during the initial phases of project development, as it facilitates a common understanding. It sets the stage for more detailed views of the system by establishing a clear framework of its external interactions and as a reference point for drilling down into more intricate process flows. A Level 0 DFD of the design ecosystem for PDT for well-being is shown in Figure 3-2.

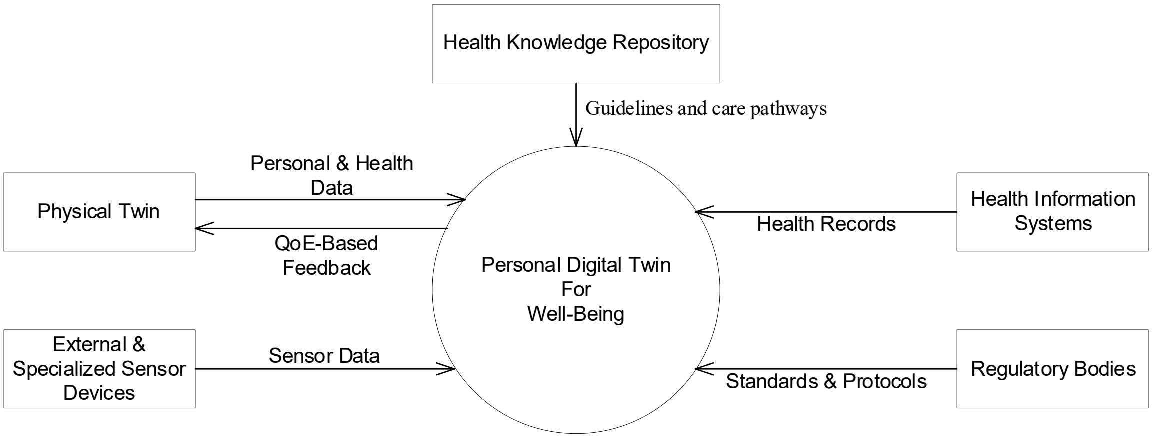


Figure 3-2. Level 0 data flow diagram of the design ecosystem for personal digital twins for well-being.

The Level 0 Data Flow Diagram captures the overarching data exchanges within the Personal Digital Twin (PDT) for the Well-Being system and to external entities. The diagram centers around the PDT, where data consumption, processing, and user interaction occur. Each data flow and process outlined contributes to the holistic functioning of the PDT, ensuring it meets its goal of enhancing the well-being of the Physical Twin (PT).

Physical Twin. This external entity comprises the real-world counterpart to the PDT from whom personal and health-related data is sourced and to whom QoE-based feedback is directed.

Health Information Systems. This entity represents the backend systems that store and manage the PT's health records, interfacing with the PDT to provide a historical and clinical context for the PT's current state.

Health Knowledge Repository. This external entity is a centralized hub that stores comprehensive healthcare protocols and well-being guidelines to inform the PDT’s recommended well-being interventions. It integrates global medical standards and current research to guide personalized health advice.

Regulatory Bodies. This entity represents the authorities mandate standards and protocols to which the system must conform, ensuring the PDT operates within the required legal and ethical frameworks.

External & Specialized Sensor Devices. This data flow represents the data collected from hardware instruments that capture real-time physiological and environmental data, providing objective metrics crucial for health monitoring within the Personal Digital Twin system.

Personal & Health Data. This data flow represents information gathered directly from the PT, encompassing health metrics, personal metrics (e.g., height and weight), and subjective well-being measures. The PDT for Well-Being system ingests this data to inform its health analysis and recommendations.

QoE-Based Feedback. The feedback loop is critical for system refinement. It is a flow of user experience data back to the PDT system derived from the PT's interactions with the system. This feedback informs adjustments and enhancements, tailoring the system to meet the PT's satisfaction and usability standards.

Sensor Data. This data flow represents raw data from external and specialized sensor devices. These devices collect physiological and environmental parameters, essential for the PDT system's continuous monitoring and health assessment processes.

Health Records. This data flow encompasses the PT's electronic health records (EHRs), retrieved from Health Information Systems. It integrates clinical histories, diagnostic results, and treatment plans, enriching PDT's data for comprehensive health management.

Standards & Protocols. A vital data flow from Regulatory Bodies includes regulations, guidelines, and interoperability standards that the PDT system adheres to, ensuring compliance and safeguarding the PT's data privacy and system security.

Together, the external entities, the system, and the data flows illustrate the nature of the design ecosystem for PDTs for Well-Being. The following section will provide a more granular view of the design ecosystem with a component diagram.

3.3. Component diagram

The Component Diagram is crucial in illustrating the structural aspects of the Design Ecosystem for PDTs for well-being. The ecosystem is broken down into its fundamental parts, detailing the interactions and dependencies among the components that make up the three modules: “Data Source,” “AI-Inference Engine,” and “Multimodal Interaction.”

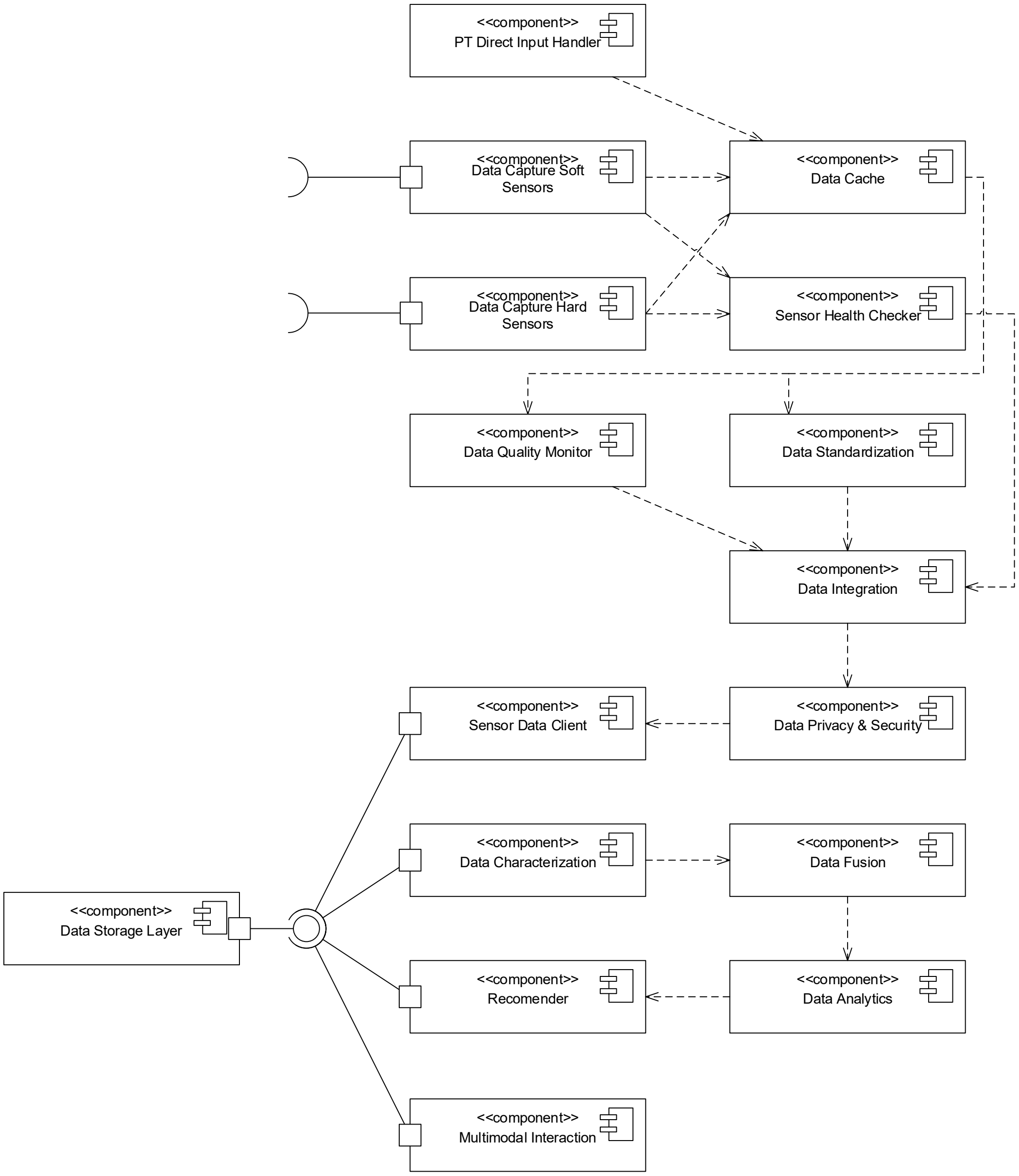


Figure 3-3. Component diagram of the design ecosystem for personal digital twins for well-being.

The component diagram delineates each component's roles and responsibilities, ensuring a cohesive operational framework. This visual representation helps grasp the technical underpinnings and workflow integration points of the PDT design ecosystem. It also aids in identifying potential areas for scalability, maintenance, and enhancements, which may be helpful when using the ecosystem. Each of the components for the design ecosystem is detailed next; the diagram itself is shown in Figure 3-3. Each of the components in the diagram from Figure 3-3 is detailed next.

PT Direct Input Handler. The PT Direct Input Handler is the interactive gateway for the Physical Twin (PT) to provide personal data directly. This component is responsible for actively soliciting and capturing PT-reported information, including responses to health surveys, symptom reporting, and other forms of direct input, such as recording headaches or nausea. This component needs a user interface that prompts the PT for specific information when needed. It helps collect a variety of inputs ranging from subjective wellness assessments to objective symptom declarations. The User Input Handler must be designed to ensure that the PT's contributions are streamlined into the system's data flow, enriching the dataset. In addition to data collection, the User Input Handler is configured to validate the inputs for consistency and completeness before they are channelled into the system for further processing. This step is necessary for maintaining the integrity of the data collected.

Data Capture Soft Sensors. This component specializes in acquiring data from non-hardware sources, such as health information systems and social networks. It acts as a collector for information generated by third-party services or applications through REST APIs or similar endpoints based on network communication (i.e. web sockets or similar), enabling efficient and automated data retrieval. This component must be capable of polling these APIs to query and subscribe to new data updates, taking advantage of webhook mechanisms or similar subscription-based protocols to receive real-time data updates. This ensures that the PDT system remains up-to-date with the latest information without constant manual input or refreshes. As it relies heavily on the availability and accessibility of external APIs, this component requires that each interfaced system exposes an endpoint that adheres to compatible data exchange standards and protocols.

Data Capture Hard Sensors. This component is tasked with collecting physiological and environmental data through a network of hardware sensors. It is meant to communicate with virtually any device capable of transmitting the sensed data, including wearable technology such as smartwatches, fitness bands, medical devices like ECG and EEG monitors, and smartphones (leveraging their integrated sensors). These types of sensors are crucial for the real-time monitoring and recording a wide array of physical health indicators. The Data Capture Hard Sensors component operates by interfacing with the sensor hardware to extract data points such as heart rate, brain activity, movement patterns, and environmental conditions like temperature and air quality. The component also captures data in a consistent format, enabling accurate and reliable input into the system. This continuous stream of detailed sensor data provides a comprehensive view of the Physical Twin's health and the environment they interact with, forming the foundation for informed health assessments and interventions. This component must handle the high-velocity data flow inherent to real-time monitoring systems, maintaining efficiency and precision. This component would also need to support multiple and simultaneous data streams from different sensors, allowing for a scalable system that benefits from collecting various types of data and physiological signals, amounting to more accurate assessments.

Data Cache. The Data Cache Component is a temporary holding area for data collected from various sources, including direct user input, soft sensors, and hard sensors. Its primary function is to buffer incoming data, providing a staging ground where information can be quickly retrieved and readied for subsequent processing or analysis. Designed for efficiency, the Data Cache offers rapid access to recent data entries, mitigating potential delays caused by direct retrieval from more persistent storage systems. The component is critical in reducing processing load and latency by caching data, ensuring the system maintains a swift response rate. This is particularly vital for near-real-time applications within the PDT ecosystem. This component is essential for maintaining data throughput in the system, especially when managing high volumes of incoming data streams. It also safeguards against data loss in the event of intermittent connectivity issues or system failures, as the cached data can be resubmitted for processing once normal operations resume. This component has dependencies from the User Input Handler, Data Capture Soft Sensors, and Data Capture Hard Sensors components.

Sensor Health Checker. The Sensor Health Checker is a comprehensive monitoring component that oversees the functionality and reliability of both hard and soft sensors within the Personal Digital Twin (PDT) ecosystem. For hard sensors, it systematically evaluates sensor functionality to ensure accurate data collection, performing regular health checks. For soft sensors, it verifies the availability and responsiveness of health information systems and social networks to provide health and social data through the corresponding APIs. This component would need to employ algorithms that can discern sensor anomalies, discrepancies in data patterns, or hardware failures. However, if an error is detected, this component will not apply corrective measures. Instead, it would embed metadata about the sensor health status on the data streams coming from the sensors. This component has dependencies from Data Capture Soft Sensors and Data Capture Hard Sensors components.

Data Quality Monitor. This component would help assess the accuracy and integrity of the data collected from both hard and soft sensors. It systematically evaluates the data against quality metrics, identifying discrepancies, incomplete entries, or anomalies. This component would need a set of predefined rules and checks to scrutinize incoming data for errors, outliers, or inconsistencies. It can also assess the completeness and validity of data sets. The purpose of this component is not to filter out data flagged with errors but to add the corresponding annotations to the metadata for each data stream. The reasoning behind this is that depending on the use case for the PDT for well-being being designed, the datasets, even with errors or with a few "good" data points, may be fed into ML models tolerant to errors, noise or incomplete data inputs. This component has a dependency on the Data Cache component.

Data Standardization. This component's purpose is to guarantee adherence to applicable standards and norms, such as the IEEE 11073[4] or the FHIR[5] standards. This component ensures all incoming data conforms to a common set of standards and formats, facilitating seamless integration and analysis. Data Standardization involves converting data into a uniform format or standard, normalizing measurement scales, and aligning disparate data types (such as dates, numerical values, and categorical data) to ensure consistency across the platform. This process enables accurate data comparison and aggregation, especially when dealing with inputs from different geographical regions or technological platforms. By standardizing data, this component significantly enhances the interoperability of a PDT for well-being with external systems and applications, supporting a broader range of data analytics capabilities. It reduces data complexity, which minimizes processing errors and improves the efficiency of data queries and analysis. This component has a dependency on the Data Cache component.

Data Integration. This component consolidates data from diverse streams to create unified and coherent data packages. It aggregates data from the User Input Handler, Hard Sensors, and Soft Sensors, which are initially processed and temporarily stored in the Data Cache and passed through the Data Standardization component; it also adds metadata about sensor health and data quality. By compiling comprehensive data packages, it ensures that all data streams are integrated smoothly. Additionally, it receives input from the Sensor Health Checker, which includes metadata about each sensor's operational status and flags any errors. The metadata

---

[4] The IEEE 11073 Personal Health Device (PHD) standards focus on ensuring the interoperability of personal health devices like weighing scales, blood pressure monitors, and blood glucose monitors. Unlike earlier IEEE 11073 standards that targeted hospital-use devices, PHD standards emphasize personal-use devices with a simpler device model. These standards are designed to facilitate open, standardized control of information exchange between personal health devices and managers such as cell phones, personal computers, and health gateways.

[5] FHIR (Fast Healthcare Interoperability Resources) is a standard describing data formats and elements (known as "resources") and an application programming interface (API) for exchanging electronic health records. FHIR is designed to enable healthcare information to be available, discoverable, and understandable globally, and to be used by computer applications easily.

generated from the Data Quality Monitor is also incorporated at this stage. This metadata is crucial for contextualizing the data, particularly in understanding potential biases or anomalies that might affect analysis due to sensor malfunctions. The Data Integration Component enhances the robustness and reliability of the data. This component depends on the Data Quality Monitor, Data Standardization and Sensor Health Checker.

Data Privacy & Security. This component is critical within the design ecosystem and is dedicated to ensuring the privacy and security of the Physical Twin's (PT) data. It implements a range of technical measures to safeguard personal and sensitive data. One example is encryption techniques, which would help ensure that only the designated PDT could make sense of it. This component is dependent on Data Integration.

Sensor Data Client. The Data Storage component is the intermediary responsible for sending data to the central data repository. It serves as the crucial link that ensures data, once cleared by Data Security, is securely and efficiently transferred to the designated storage system, functioning effectively as a client to the database. Upon receiving data processed and secured by the Data Security, this component orchestrates the storage operations, encapsulating the complexity of data handling. This component utility is more evident when the data repository is hosted remotely on the cloud or an external server. Otherwise, the functionality of this component is rather simple. When the data repository is hosted externally, this component manages the serialization and transmission of data, ensuring that it is packaged and sent in formats that are compatible with the data repository. The Data Storage component also handles connection management, maintains a stable and secure link to the data repository, and manages retry logic and handling errors to cope with network failures or data transmission errors. This component is built to ensure robustness in data handling, providing mechanisms such as batching, caching, and transaction management to enhance data throughput and integrity during storage operations. By serving as the client to the data repository, the Data Storage component plays a pivotal role in the data lifecycle management within the PDT ecosystem, ensuring that data is stored securely and ready for quick retrieval and analysis as needed. This component depends on Data Security; it also has a required interface to write data into the data repository.

Data Storage Layer. This component is the foundational component within the design ecosystem for PDTs for well-being, functioning as an abstract representation of the data repository or database. This component is responsible for the persistent storage of all collected and processed sensor data, ensuring it is securely housed locally or in a cloud-based environment. The Data Storage Layer encapsulates the complexities of data storage management, providing a uniform interface for storing diverse data types from various sources within the design ecosystem. It supports a range of storage operations, including data insertion, retrieval, update, and deletion, facilitating efficient data management practices. By abstracting the specifics of data storage, this component allows the PDT system to interact uniformly with different storage technologies without being tied to any implementation. It manages the storage, retrieval, updating, and deletion of data in a way that is transparent to the rest of the system, ensuring smooth data flows and integrity across different platforms and devices. The Data Storage Layer Component is designed to provide a flexible and scalable storage solution that supports the high demands of data-intensive operations typical in PDT environments. It ensures data consistency and durability, efficiently handling large volumes of data from various sources. Designed to handle high volumes of data, the Data Storage Layer ensures data integrity, consistency, and availability. This abstraction simplifies the system architecture and enhances its adaptability to integrate with new storage technologies as they emerge. This component has a provided interface allowing data insertion and recovery from other components.

Data Characterization. This component is designed to process and refine the data collected and pre-processed from various sources compiled during Data Integration. It focuses on summarizing and extracting key parameters and properties used to assess the well-being of the Physical Twin (PT). This component's sole purpose is to extract features and characteristics from the data collected from the sensors. The component may perform processes like feature extraction and data augmentation where appropriate. This component has a required interface to recover data from the Data Storage Layer.

Data Fusion. This component assembles the input datasets for the ML models in the Data Analytics component. Each ML model may accept different features as input from the Data Characterization component. For example, a classifier may only take a subset of channels of the ECG data stream as input. In contrast, another ML model may input the whole set of ECG channels plus laboratory results. This component would fuse the necessary data streams to form the corresponding inputs for each ML model. This component is dependent on Data Characterization.

Data Analytics. This component runs Machine Learning models using data from Data Fusion as input. This module is crucial for extracting meaningful insights from the data, employing machine learning techniques to detect health issues, predict outcomes, and support decision-making that may prompt lifestyle changes. This component is dependent on Data Fusion.

Recommender. This component is a critical part of the Personal Digital Twin (PDT) ecosystem, functioning as the conduit through which personalized recommendations for well-being management and improvement are formulated. By leveraging insights derived from Data Analytics, this component continuously monitors the health state of the PT; it can also identify potential health problems; in turn, this component crafts tailored suggestions to enhance the PT's well-being or prevent future health issues. The recommendations must align with the PT's unique preferences and contextual realities. Whether suggesting lifestyle adjustments, medical interventions, or behavioural changes, the recommendations are designed to resonate with and be actionable for the PT. By providing contextually appropriate recommendations, the Recommender aims to improve health outcomes and enhance the PT's engagement, adherence and compliance with suggested actions. This approach contributes to ensuring that a PDT for well-being delivers on its promise of positively impacting an individual’s health. This component depends on Data Analytics and has a required interface, i.e., a way to send the recommendations to the Data Storage Layer.

Multimodal Interaction. This component acts as a bridge through which the PDT communicates with the PT by rendering a digital representation that includes health status, analytical inferences, and personalized recommendations generated in the Recommender. Central to this component is its role in the QoE-driven feedback loop, which emphasizes optimizing the PT's overall experience with the PDT system. It dynamically adjusts the interaction based on real-time user feedback and changes in the PT's health data, ensuring that the system remains aligned with the PT's preferences and situational demands. When the PT adopts recommendations presented through this component, changes in the PT’s health will start to manifest (either improving or deteriorating), which will be picked up by the PDT when collecting new data and adjusting recommendations accordingly, thus closing the feedback loop. This component leverages a variety of interaction modalities to enhance user engagement and ensure effective communication. These modalities include visual displays, auditory feedback, and even virtual or augmented reality environments, each chosen based on the context of interaction and the specific needs of the PT. This adaptive interaction strategy improves the PT’s engagement with the system. It enhances the effectiveness of the health management strategies recommended by the PDT, promoting better health outcomes and greater satisfaction with the system. This component has a required interface to recover the recommendations generated by the Recommender from the Data Storage Layer.

The structural elements and interrelationships of the design ecosystem for PDTs for well-being have been delineated with the component diagram. The next diagram, A Level 2 Data Flow Diagram (L2 DFD), is useful to address the operational dynamics of the design ecosystem. The L2 DFD extends the foundational understanding provided by the Component Diagram by illustrating how data moves through the system, detailing the processes that transform raw inputs into meaningful outputs. While the Component Diagram effectively sets the stage by outlining the structural elements of the ecosystem, a scaffold of sorts, for the design ecosystem, the L2 DFD offers a view of the data interactions, pinpointing where and how data is processed, stored, and transmitted.

3.4. Level 2 data flow diagram

A Level 2 Data Flow Diagram (DFD) provides a more detailed view of a system's processes than its Level 0 and Level 1 counterparts. Level 2 DFD presents the system’s processes in a more specific, granular way, offering a deeper insight into the operations of a particular subsystem or function within the overall system architecture. After defining the component diagram, a Level 1 DFD could be redundant as identifying the processes on the Level 2 DFD level was a small step from the component diagram, thus a Level 1 DFD would not offer much value to the discussion of the design ecosystem. Here are the key elements of a Level 2 DFD:

- Processes: A process represents a specific, detailed function or operation within the system that transforms input data into output data. Each process is depicted as a distinct element that performs a well-defined task, contributing to the overall functionality of the system. Processes in a Level 2 DFD are typically more granular and detailed compared to those in higher-level diagrams, focusing on the internal mechanics and sub-tasks of a larger process identified in a Level 1 DFD.
- Data Stores: Level 2 DFDs introduce data stores that are relevant to the specific sub-processes being examined. Data stores represent where data rests within the system, such as databases or temporary storage areas, and they are depicted as open-ended rectangles or lines.
- External Entities: These remain consistent with higher-level DFDs but are connected specifically to the processes where external interactions occur. External entities are sources or destinations of data outside the system boundary.

- Data Flows: Arrows depict the flow of data between processes, external entities, and data stores. These flows detail the inputs and outputs of each process, illustrating how data is manipulated and transferred through the system.

The L2 DFD diagram traces the data lifecycle from initial collection through processing stages and eventual utilization. Since its conception, the design ecosystem has favoured adopting and supporting standards and protocols when interacting with entities like health information systems and regulatory bodies and throughout the data life cycle.

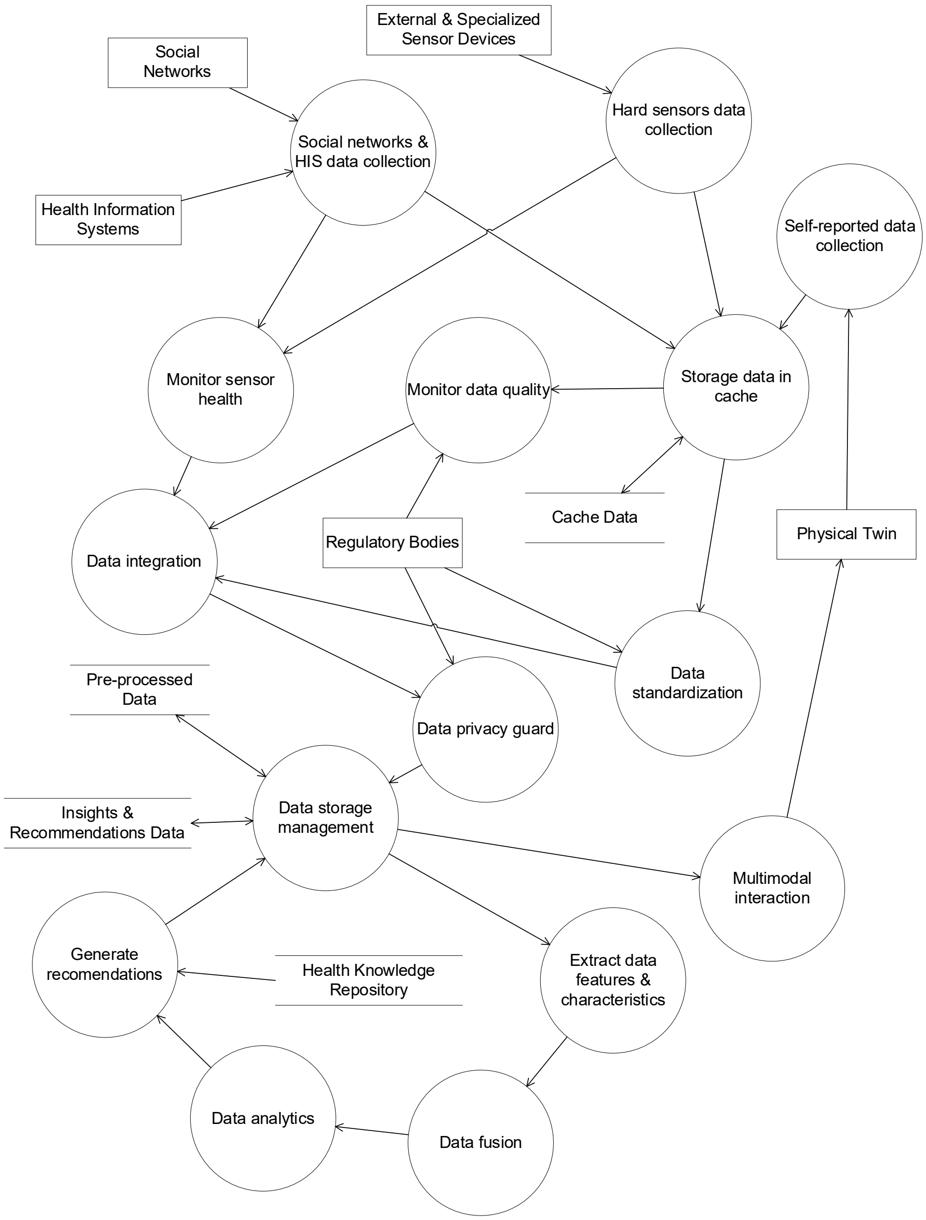


Figure 3-4. Level 2 data flow diagram of the design ecosystem for personal digital twins for well-being.

Including a Level 2 DFD mainly serves two purposes: it enhances the technical discussion on the design ecosystem and validates design decisions.

Within the DFD, External Entities are defined as systems, users, or organizations that provide or receive data, forming the interface where the PDT meets the external world. Depicted as labelled rectangles, these entities may be the data's origin and destination points. The L2 DFD from Figure 3-4 presents a few examples of external entities; Health Information Systems and

Physical Twin are examples of these. The Processes are illustrated as numbered ovals or circles; these are the transformative steps where data is processed. The Data Stores, shown as parallel lines or open rectangles, represent the repositories where this data is stored or recovered by ongoing system processes.

Together, these elements paint a picture of the PDT's data dynamics, which is necessary for understanding the practical applications of the proposed design ecosystem. They also facilitate a description of the internal workings of a PDT for well-being, enabling a seamless transition from conceptual design to tangible implementation. Figure 3-4 presents a level 2 data flow diagram of the design ecosystem for a personal digital twin for well-being.

As shown in Figure 3-4, fifteen processes conform to the design ecosystem from a data flow point of view, all detailed next, along with the data stores and external entities. The diagram also has five external entities: Health Information Systems, Social Networks, External & Specialized Sensor Devices, Physical Twin and Regulatory Bodies. The data stores identified in the diagram are Pre-Processed Data and Insights and Recommendations Data. As shown in the L2 DFD from Figure 3-4, the data originates in the external entities; then, it flows through the processes following the arrows in the diagram until the data flows back to the Physical Twin in the form of recommendations. The diagram explains each one of the processes, data stores, and external entities in detail.

Health Information Systems. This entity encompasses a comprehensive array of electronic health records and databases maintained by healthcare providers. This entity represents an authoritative source of historical and current medical data, offering a wealth of patient information, from diagnostic reports to treatment histories and recovery outcomes. In the PDT design ecosystem, Health Information Systems interface with the system through the Social networks & HIS data collection process, which facilitates the incorporation of extensive health-related data into the PDT's analytical framework. This integration ensures that the PDT can access clinical context for generating accurate health assessments and personalized wellness recommendations for the Physical Twin.

Social Networks. This entity represents social platforms that mediate social interactions between the PT and other persons. These networks provide invaluable data that captures the social aspects of the Physical Twin's life, such as interaction patterns, mood changes, and public health-related behaviours. The Social networks & HIS data collection process taps into this rich vein of qualitative data, enabling the PDT to analyze the impact of social factors on the Physical Twin's overall well-being. Integrating social media data enriches the system's understanding of the user's mental and emotional health, allowing for more holistic health insights and tailored recommendations within the digital twin framework.

External & Specialized Sensor Devices. This entity symbolizes the myriad of hardware instruments and technologies that capture real-time physiological and environmental data; the design ecosystem summarizes these into six groups, namely Fitness Trackers, Specialized Medical Devices, Electroencephalogram devices, Electrocardiogram devices, Smartphone-embedded sensors and environmental sensors. Each group may comprise sensors from different vendors and models, resulting in many possibilities. These devices play a pivotal role by feeding the Hard sensor's data collection process with precise and continuous data streams. These sensor devices are integral to the ecosystem, providing the objective metrics for comprehensive health monitoring and assessment.

Physical Twin. This entity represents the individual at the heart of the Personal Digital Twin system. It is the primary source of subjective health information, lifestyle choices, and personal preferences, contributing to the Self-reported data collection process. The physical twin's active participation is essential, as it contributes to the PDT, reflecting the unique characteristics and needs of the individual. The data provided by the Physical Twin enriches the system's understanding, enabling it to offer bespoke advice and interventions to enhance the PT's well-being.

Social networks and HIS data collection. This process represents the systematic collection of data from social media platforms and Health Information Systems. This single process manages both data sources because the way to collect data from both is essentially the same: either a call to a REST API or similar; if not, the data transfer can happen via socket-mediated communication. This dual-faceted process is designed to harness the rich, contextual data from the Physical Twin's interactions on social networks and the structured clinical data from electronic health records.

Hard sensors data collection. During this process, the PDT collects data from various hardware-based sensing devices. This includes wearable technology like fitness bands, which track physical activity and vital signs, and specialized medical devices that measure specific health parameters. This process also collects data from other devices like smartphones, more precisely, the sensors embedded in them and environmental sensors as well. This process builds a foundational dataset for the Personal Digital Twin (PDT) to analyze the Physical Twin's health status objectively.

Self-reported data collection. This process captures subjective health information directly from the Physical Twin. This process involves gathering personal health data, such as symptoms, emotional state, and lifestyle details, provided voluntarily by the user. The data might include inputs from health surveys, symptom diaries, or direct user feedback entered through visual interfaces built into the PDT system. This self-reported information complements the objective data from hard sensors, providing a holistic view of the Physical Twin's well-being. It is critical for personalizing the PDT experience, ensuring that recommendations and insights reflect the individual's subjective health experience and physiological state.

Storage data in cache. This process is responsible for mediating the temporary storage of recent data for swift access and immediate processing. This process acts as a high-speed data access layer that stores a subset of data collected during Hard sensors data collection, Social networks & HIS data collection and Self-reported data collection processes transiently. Its primary role is to enhance system performance by reducing data access latency, supporting real-time processing, and providing quick retrieval for the most frequently requested data sets. By maintaining this intermediate storage layer, the PDT system ensures that ongoing analysis and interactions with the Physical Twin are efficient and responsive to current data inputs.

Cache Data. This data store temporarily holds incoming sensor data. Its primary function is to optimize the data retrieval process. By storing this data temporarily, the Data Cache serves as a buffer that reduces latency and improves the efficiency of data processing for real-time analysis. In terms of data management, this data store implements policies for data lifecycle, such as time-to-live (TTL) parameters, to ensure that only relevant data is kept at hand while outdated information is purged. This selective retention policy conserves resources and aligns with data privacy guardrails by minimizing the exposure window for sensitive information.

Monitor sensor health. This process comprises assessing the operational integrity of hard and soft sensors involved in data collection. It involves continuously checking the status and functionality of these sensors, verifying that they provide accurate and reliable data. It is crucial for detecting potential issues such as sensor malfunctions, data corruption, or signal loss. The assessments produced during this process are integrated into the data stream to flag issues with the data as it is being collected.

Regulatory Bodies. This external entity represents governmental or authoritative organizations that set standards and regulations governing health data management, transfer, and use. This entity provides crucial input to Monitor data quality, Data privacy guard and data standardization processes for ensuring that the PDT adheres to legal requirements, privacy policies, and ethical guidelines, particularly in the handling, storing, and processing of sensitive health information. Interaction with Regulatory Bodies typically involves compliance checks and updates to operational protocols to align with new or revised regulations. This continuous engagement helps maintain the system’s credibility and trustworthiness, safeguarding the interests and privacy of the Physical Twin.

Monitor data quality. This process involves systematic checks to ensure that the incoming data from hard and soft sensors and self-reported inputs meet established quality standards. It includes validating the data's accuracy, completeness, consistency, and timeliness to prevent errors and discrepancies that could impact the system's analyses and recommendations. Anomalies or issues are identified and flagged early in the data flow during this process.

Data standardization. All incoming data, regardless of its source, is converted into a consistent format suitable for integration and analysis within PDT. This process involves normalizing diverse data types—from sensor readings to self-reported information—into a uniform structure and aligning measurement scales. By standardizing data, this process facilitates seamless data integration. It enhances the accuracy and efficiency of subsequent analytical operations within the PDT, which is crucial for generating reliable health insights and recommendations for the Physical Twin.

Data integration. This process combines data from various sources, including sensor data, self-reported information, and health records, into a coherent dataset. This process ensures that all the standardized data elements are merged accurately, maintaining their relationships and contexts. The integrated data is then prepared for further analysis and processing, forming a comprehensive view that supports the Personal Digital Twin (PDT) system in delivering personalized insights and recommendations for the Physical Twin’s health and well-being.

Data privacy guard. This process takes input from regulatory bodies that implement pertinent privacy and security measures for the data collected from the sensors in the PDT. The measures implemented and applied during this process include but are not limited to, encryption, removing full addresses, removing full names, and regulatory compliance checks. It effectively shields sensitive information from potential breaches, safeguarding the integrity of the data as it moves through the system.

Data storage management. This process oversees the PDT's organization, retention, and data maintenance. It ensures that data is securely stored in a structured manner, facilitating efficient retrieval and updates. It manages the long-term storage of historical data and provides data for ongoing operational needs, ensuring data is available and consistent within the PDT.

Pre-processed Data. This is a data store; it serves as a repository for data that has undergone initial pre-processing during the standardization. This store ensures that data remains readily accessible for further refinement and processing steps within the Personal Digital Twin (PDT) system, maintaining its readiness.

Insights and Recommendations Data. This data store holds the finalized outputs from the data analytics processes, specifically tailored insights and actionable recommendations for the Physical Twin. It is essential to retain the valuable conclusions derived from the Personal Digital Twin (PDT) system's analysis and ensure these insights are available for prompt delivery to the Physical Twin through various interaction modalities, thus supporting informed decision-making and personalized health management.

Extract data features. This process is responsible for identifying and isolating relevant features from the integrated data set compiled with the data collected and pre-processed during the initial steps (i.e. Pre-Processed Data). It implements algorithms designed to extract characteristics and features necessary to use as inputs to obtain insights and assess the PT's health state. These extraction algorithms may be used to calculate features and characteristics that may include but are not limited to digital signal processing, digital image processing, knowledge transfer, data augmentation or dimensionality reduction to distill the data into a more suitable form for effective analysis.

Data fusion. This process combines the features generated during the extract data features process. Multiple data streams and sources are used to extract features. At the same time, the features used as inputs to discover insights on the PT's health status may originate from more than one source or data stream. During Data Fusion, the features originating from different sources and data streams may be combined, rearranged, and consolidated differently to become inputs for the data analytics process.

Data analytics. During this process, advanced analytical techniques are applied to the fused data to uncover patterns, derive insights, and predict PT's health status, trends, and health risk assessments. This process leverages statistical models, machine learning algorithms, and other data science tools to interpret complex data sets, providing the basis for actionable health recommendations and strategic interventions for the Physical Twin.

Generate recommendations. During this process, the PDT formulates specific, actionable recommendations for the Physical Twin based on the analyzed data, ensuring that the recommendations are tailored to the Physical Twin's current health status and lifestyle. This process uses the Health Knowledge Repository to formulate pertinent, safe, and beneficial recommendations for the PT.

Health Knowledge Repository. This external entity is a central resource that compiles a spectrum of evidence-based treatment guidelines, care pathways, and well-being interventions in any or all of the three aspects of health, namely physical, mental and social. This entity is necessary for the Generate Recommendations process, supplying updated medical and well-being information to ensure recommendations are pertinent, safe, beneficial and reflective of the latest best practices.

Multimodal interaction. This process manages the delivery of the generated recommendations and health insights to the Physical Twin using various modalities of interaction, including, but not limited to, video, audio, and a hybrid like VR, to ensure that the information is accessible and understandable, thereby enhancing user engagement and the effectiveness of the interventions.

The Level 2 Data Flow Diagram (L2 DFD) explained the data lifecycle within the PDT for the well-being design ecosystem, mapping out the intricate flow from collection through various transformation stages to the final actionable output. By encapsulating the roles of external entities, detailed processes, and data stores, the L2 DFD serves as a technical cornerstone for understanding the data dynamics within a PDT for well-being and as a crucial validation tool for design decisions. With the careful delineation of each component's expected behaviour, researchers may use it as a guide to model the interactions and dependencies when implementing a PDT for well-being.

3.5. Use case diagram

A Use Case diagram facilitates an exploration of the system's functionality and the user's interaction with it, offering a user-centered view (i.e. the physical twin) that complements the data-centric approach of the DFD. The next section will present a use case diagram of the PDT for the well-being design ecosystem. In software design and system analysis, use case diagrams are instrumental in capturing and illustrating the functional requirements of a system from the perspective of the under. A Use Case diagram, as defined in the Unified Modeling Language (UML) standards, is a representation that outlines the system's interactions with external agents – the actors – and delineates the services, features, and operations it provides. It is a high-level depiction of the system's functionalities and environment, highlighting the relationships between the user (i.e. actor) and the system. This section will introduce the Use Case diagram for the PDT for the well-being design ecosystem.

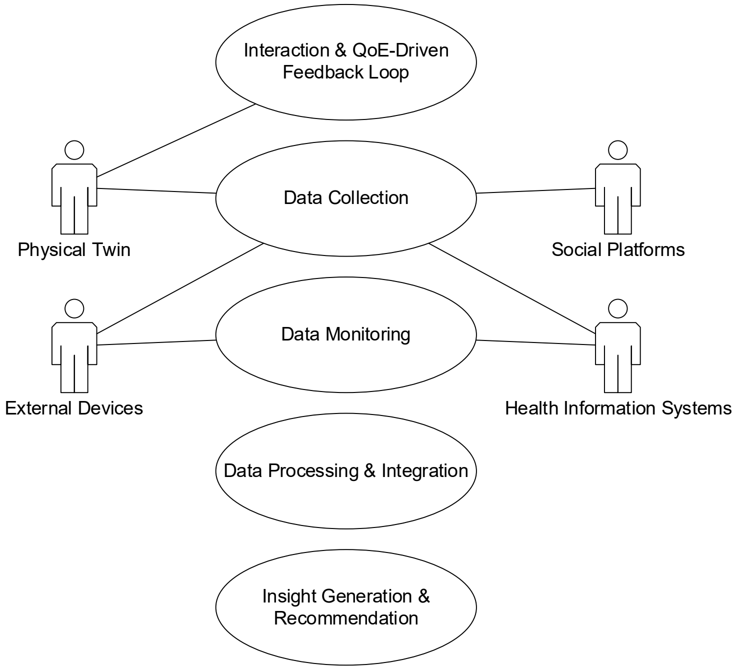


Figure 3-5. Use-case diagram of the design ecosystem for personal digital twins for well-being.

Figure 3-5 shows a use-case diagram that outlines the functional requirements for a PDT for well-being. This diagram is also part of the design ecosystem. However, two of the use cases show no direct association as these would need to be associated with the “system”, in this case, the PDT itself. The reasoning for this is that the PDT has an association with all five use cases, and drawing it would just add unnecessary complexity to the diagram alone; thus, the corresponding associations between the PDT and the five use cases are obviated in favour of a leaner diagram. Next, each of the use cases is detailed.

Data Collection

- Actors: Physical Twin, External Devices, Social Platforms, Health Information Systems
- Description: This use case encompasses all data collection mechanisms. It includes collecting physical data via sensors, self-reported data from the Physical Twin, social interaction data from social platforms, and health and clinical data from Health Information Systems. The use case will be split into sub-use cases for each data source to maintain clarity.

Data Maintenance

- Actors: External Devices
- Description: Data Maintenance includes maintaining the health status of sensors and ensuring the data quality and integrity from all collection points. This involves regular diagnostics, calibration, and validation processes.

Data Processing and Integration

- Actors: None (System Process)
- Description: This system takes the collected data, standardizes it, and integrates it into a unified model. This data is then ready for analysis. It's a high-level function that describes the necessary steps to prepare data for insights generation.

Insight Generation and Recommendation

- Actors: None (System Process)
- Description: Another system-level function that describes the analysis of integrated data to generate health insights and, based on these insights, crafting personalized well-being recommendations for the Physical Twin.

Insight and Recommendation Delivery

- Actors: None (System Process)
- Description: This represents how the generated insights and recommendations are delivered to the Physical Twin. This can include various modalities, such as visual, auditory, or tactile feedback systems.

Interaction and Feedback Loop

- Actors: Physical Twin
- Description: This covers the Physical Twin’s interactions with the PDT system, providing feedback on recommendations, updating personal data, and modifying preferences or settings. It's essential for the adaptability and personalization of the PDT system.

This diagram serves as a guidepost, anchoring the system's functional requirements to real-world scenarios where the interactions of users, technology, and data converge. It elucidates the pathways through which stakeholders engage with the system, showcasing a comprehensive suite of actions from data input to the receipt of personalized health insights. The use cases detailed herein form a blueprint for designers and developers, encapsulating the essence of user interaction and system responsiveness. They are the narratives that, when woven together, create a cohesive and user-centric experience—ensuring that the PDT ecosystem is not only technologically sound but also intuitively aligned with the needs and well-being of its users. Moving past this section, the thesis delves into the coordination of the different elements of a PDT for well-being designed using the design ecosystem (TMW). This diagram will answer when the other elements go into action.

3.6. State Machine Diagram

A state machine diagram represents the states and transitions of a system. It will help visualize the states in which a PDT for well-being can be in, as well as how it transitions from one state to another based on events, conditions, or activities. State machine diagrams are particularly useful for modeling the behavior of classes, interfaces, or collaborations, especially when the systems involve complex logic or have numerous possible states that depend on events or conditions.

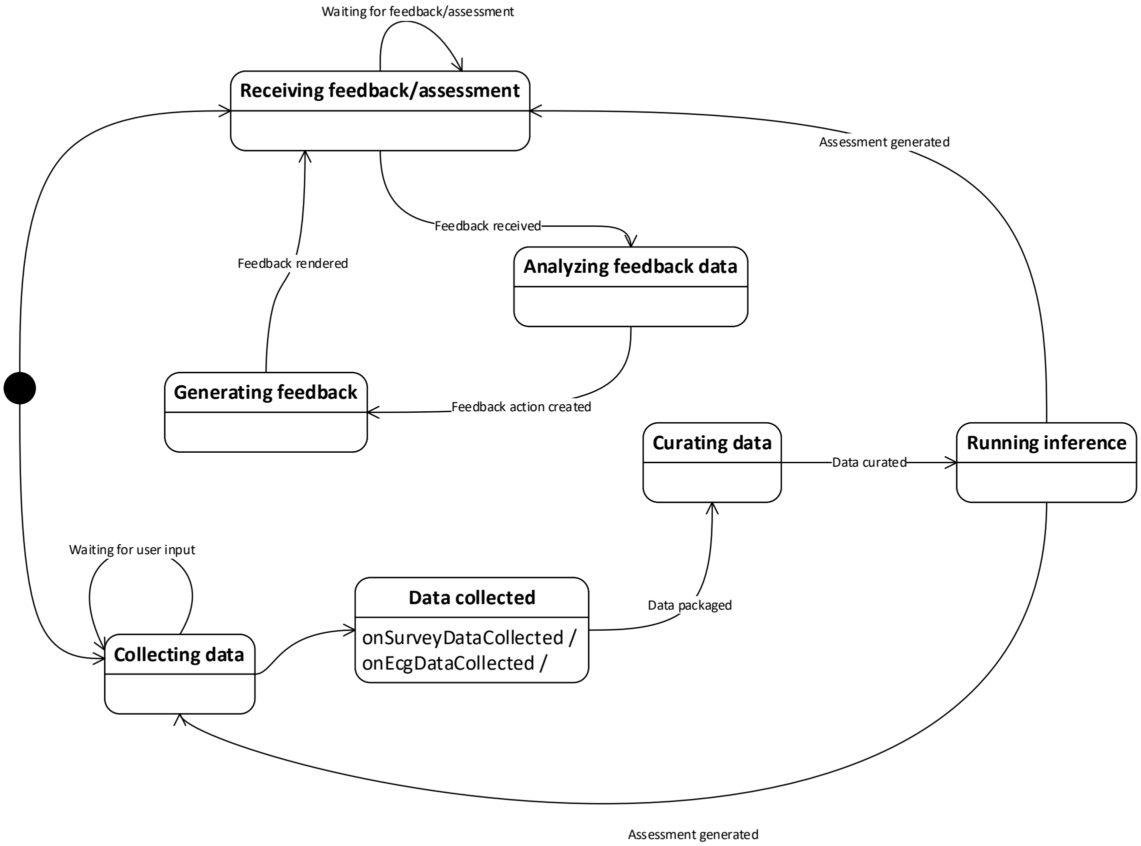


Figure 3-6. State machine diagram of the design ecosystem for personal digital twins for well-being.

The states are depicted as rectangles with rounded corners, transitions are the arrows connecting states, and the conditions that trigger these transitions are described next to the transitions.

Figure 3-6 shows a state machine diagram that can be useful for understanding the dynamics and transitions in the PDT for well-being.

Collecting Data. The system waits for the user to input data or for the sensor to read data directly from the user. When data is collected (survey or ECG data), the system transitions to the next state, “Data Collected.”

Data Collected. Once data is collected, the system packages or encapsulates it in a way that makes sense for the next state, “Data Curation.”

Curating Data. In this state, the system cleans and applies transformations to the data as the next state requires. Once the data is curated, it is stored in local storage, and the system transitions to the next state, “Running inference.”

Running Inference. In this state, the system selects the ML model and algorithm to feed the curated data depending on the type of curated data available (e.g., Survey or ECG) and also depending on the expected inference (sometimes the same curated data may be used for different purposes; for example, beats per minute and heart rate variation could be extracted from the same ECG signal). Once an assessment is generated, two parallel states are triggered: "Sending Assessment” and “Receiving feedback/assessment.”

Receiving Feedback/Assessment. In this state, the system constantly listens for new assessments produced during the “Running Inference” state. Once a new assessment is received, the system transitions to “Analyzing feedback data.”

Analyzing Feedback Data. In this state, the system analyzes the feedback/assessment from the previous state, and an action is decided depending on the nature of the feedback/assessment. Then, the system transitions to “Generating feedback”.

Generating Feedback. In this state, the system renders the user's feedback based on the analysis from the previous state.

Chapters 4 and 5 present the design and implementation of two PDTs for well-being, “COVID Me” and “Cardio Twin” respectively. The proposed design ecosystem was used in the design of both PDTs.

Chapter 4

4. COVIDMe

Testing was of the utmost importance in keeping COVID-19 from spreading extensively; when COVID-19 was declared a pandemic and a crisis in 2020, health authorities around the globe implemented aggressive testing campaigns to get a more accurate estimation of the number of infected people as part of their strategies to stop the spread of the virus [61], [62], [63]. After all, how can any government help people without knowing who is infected? During the declared crisis there were two types of tests to detect current infection with SARS-CoV-2 (the virus causing the COVID-19 disease), the antigen test and the Nucleic Acid Amplification Test (NAATs); the latter considered by the Centers for Disease Control and Prevention (CDC) as the most sensitive tests [64], [65], [66]. The reverse transcription-polymerase chain reaction (RT-PCR) test is likely the most popular among the NAATs family. During the crisis, unfortunately, testing rates in developing countries such as Mexico were significantly lower than in developed countries; the costs associated with the RT-PCR tests were probably a factor. However, the fact remains that as of November 2021, Mexico reported having performed under 90 thousand tests per 1 Million people, while countries such as Germany, Canada, the UK, the USA and others reported close to 1 Million tests per 1 Million people and above; around the same time in November 2021, the country with the highest testing rate among the countries most affected by the pandemic was the UK with over 5 Million tests per 1 Million population [67]. As of November 2021, if every person in the UK got tested for COVID-19, each person would have been tested for COVID-19 close to 5 times on average. The UK at the time (i.e. November 2021) was among the best-performing countries in testing. Yet, health authorities in that country would have had a problem figuring out the exact number of infected people at any given time because not everyone was being tested regularly, asymptomatic and people who experienced mild symptoms of COVID-19 would have been very hard to detect. The Percent Positive (PP) is a metric designed to assess community transmission of the disease and precisely overcome these challenges. The PP is the percentage of positive results from the total tests achieved inside a health jurisdiction. The interpretation of PP may vary depending on the circumstances of the jurisdiction being monitored. A high PP in a jurisdiction with low testing rates may indicate that more testing is needed. In contrast, a high PP in a jurisdiction with a high testing rate may indicate high levels of community transmission [68]. Sadly enough, some developing countries at the time presented high PP rates, in some cases 20% and above, in combination with significantly lower testing rates when compared to developed countries [69].

As explained, a higher testing rate is preferred. However, common sense helps us understand that health authorities may not be able to test 100% of the population periodically due to budgetary limitations and material limitations such as a shortage of personnel. In lieu of that, COVIDMe would be helpful to address the issue of continuous COVID-19 testing by capturing the health state of the PTs and screening individuals for COVID-19 indirectly in the process. In turn, the collective data from each COVIDMe instance may indirectly provide insight into the progress of COVID-19 infections by leveraging the relationships between the PTs. In this chapter, we'll discuss the implementation of COVIDMe to validate the design ecosystem for a PDT for well-being presented in Chapter 3. Secondly, given that COVIDMe is a tool with the potential to detect COVID-19 infections automatically and the importance of early detection of COVID-19, the chapter also provides an insight into the effectiveness and accuracy of the ML model, which in turn serves as evidence to back the use of analytics in automated well-being assessment and diagnostics on the PT's health.

4.1. Automatic Detection of COVID-19

The effectiveness of COVIDMe in assessing the current health status of the PT relies mainly on the accuracy of the COVID-19 classification component, which has a machine learning classifier at its core designed explicitly for this purpose (detect COVID-19). Researchers around the globe adopted several strategies to develop and train classifiers in detecting COVID-19; next, some examples are presented; the type of data used to train these classifiers includes images, sound or signal-based, text and combinations of the previous.

A convolutional neural network (CNN) is a particular type of neural network initially introduced by LeCun et al. in [70]. Convolutional neural networks have proven to be exceptionally effective when used to tackle tasks involving image recognition; it's not a surprise that teams all over the globe turn to these networks to train classifiers to detect the presence of CVOVID in Computer-Aided Tomography (CAT) Scans, conventional X-ray Scans, Ultrasound Scans or any other form of medical imaging technique. Many works focus on medical imaging, primarily scans from the chest area of individuals suspected of being infected with the virus and healthy people. The argument to justify the focus on chest scans is that the SARS-CoV-2 virus creates lesions in the lungs of affected persons (primarily in severe cases). According to the task for which the algorithms were trained, these may be grouped into segmentation and direct classification; the difference being that for the first, the classifier identifies COVID-19 lesions and for the later, the classifier identifies the presence of COVID-19 without properly delimiting lesions within the images. Examples of image-based classifiers are presented in [71], [72], [73], [74], [75], [76]. With the image approach, the classifiers present a baseline accuracy of 0.8. In some cases, a potential risk of overfitting a dataset was identified, most likely due to small datasets used for training and testing, which is not ideal for implementing deep learning or CNNs.

Although less popular, another approach to automatically detecting COVID-19 by applying machine learning is using sounds like coughing or abnormal breathing (e.g., whizzing sounds) captured with digital stethoscopes. In [77] Dash et al. present a classifier to identify possible viral infections by extracting cepstral features from coughing and breathing sounds. The classifier takes the cepstral features to train a Support Vector Machine (SVM) classifier and achieves an accuracy of 0.85. However, the classifier reported in this work was not trained to classify COVID-19 specifically. Instead, it is trained to detect flu-like viral infections.

In other instances, classifiers are trained with features extracted from cough sounds to detect the presence of COVID-19 [78], [79] .

Yet another approach is using data extracted directly from medical laboratory tests (i.e., blood tests). The results with this kind of data are promising, and it seems that the features obtained from the lab tests are well suited to be used with more conventional machine learning algorithms (not deep learning) [80], [81], [82], [83].

COVIDMe took a different path for the Inference module and used self-reported symptoms as features to assess the Physical Twin's health. It is acknowledged that other types of data (e.g., image-based or blood-based) may lead to more accurate predictions. However, COVIDMe requires frequent data collection from the PT, making it possible to use other data types, such as CAT Scans, X-ray scans, or medical laboratory test results.

4.2. The design of the COVIDMe platform

COVIDMe draws directly from the design ecosystem presented in Chapter 3. However, it was decided to streamline the design because COVIDMe is only concerned with copying a single aspect of the human body: the health status concerning COVID-19. The resulting platform also has three primary modules: “Data Source,” “Inference Engine,” and “Feedback.”

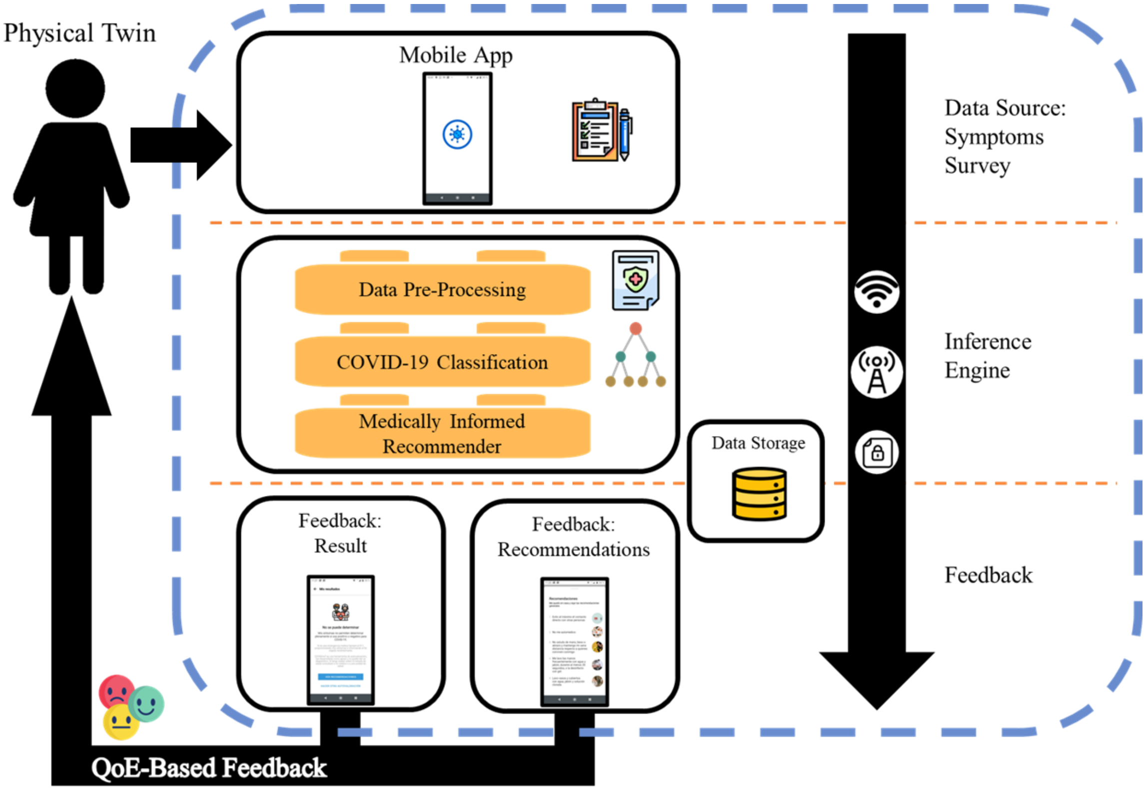
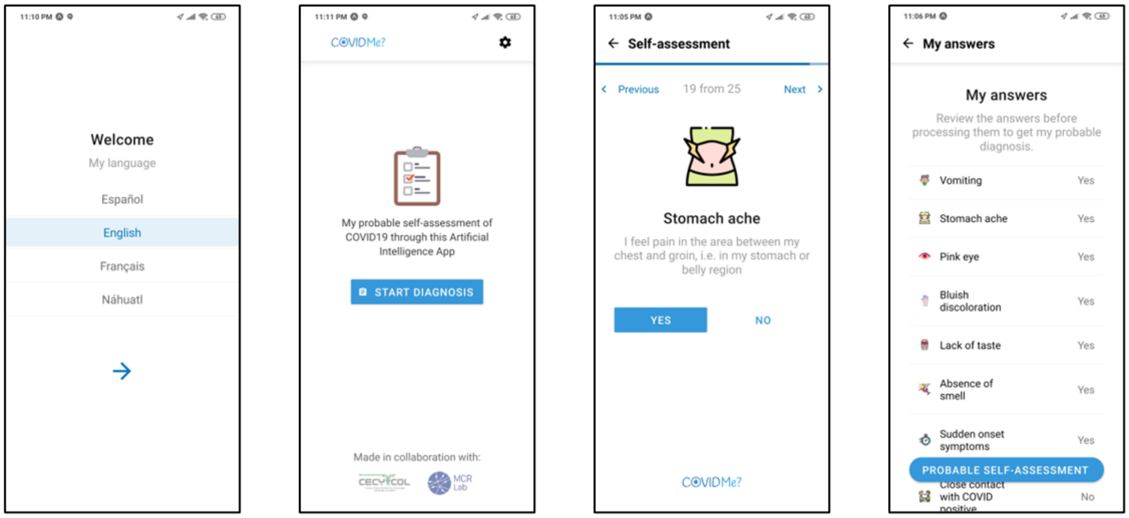


Figure 4-1. COVIDMe, a personal digital twin for health in the context of COVID-19.

Figure 4-1 shows the final approach to COVIDMe, which creates a digital copy of a small portion of the PT, more precisely, a set of symptoms required for the classifier to predict a chance for the PT to be infected with SARS-CoV-2 (COVID-19). The classifier presented in [84] was used to implement the “COVID-19 Classification” component in the Inference Engine from Figure 4-1. The classifier required an array of true/false values representing the presence of symptoms associated with COVID-19, some of which are more directly associated with COVID-19 or a respiratory infection than others; the complete list is fever, cough, sudden onset of symptoms, known covid contact, odynophagia, rhinorrhea, irritability, chest pain, cephalea, chills, diarrhea, fatigue loss appetite, dyspnea, myalgia, arthralgia, conjunctivitis, abdominal pain, polypnea, vomit, anosmia, dysgeusia, cyanosis.



My probable self-assessment of COVID19 through this Artificial Intelligence App

Made in collaboration with:

Stomach ache

I feel pain in the area between my chest and groin, i.e. in my stomach or belly region

My answers

Review the answers before processing them to get my probable diagnosis.

Vomiting	Yes
Stomach ache	Yes
Pink eye	Yes
Bluish discoloration	Yes
Lack of taste	Yes
Absence of smell	Yes
Sudden onset symptoms	Yes
Close contact with COVID positive	No

a. COVIDMe:
Select language

b. COVIDMe:
Start survey

c. COVIDMe:
A survey question

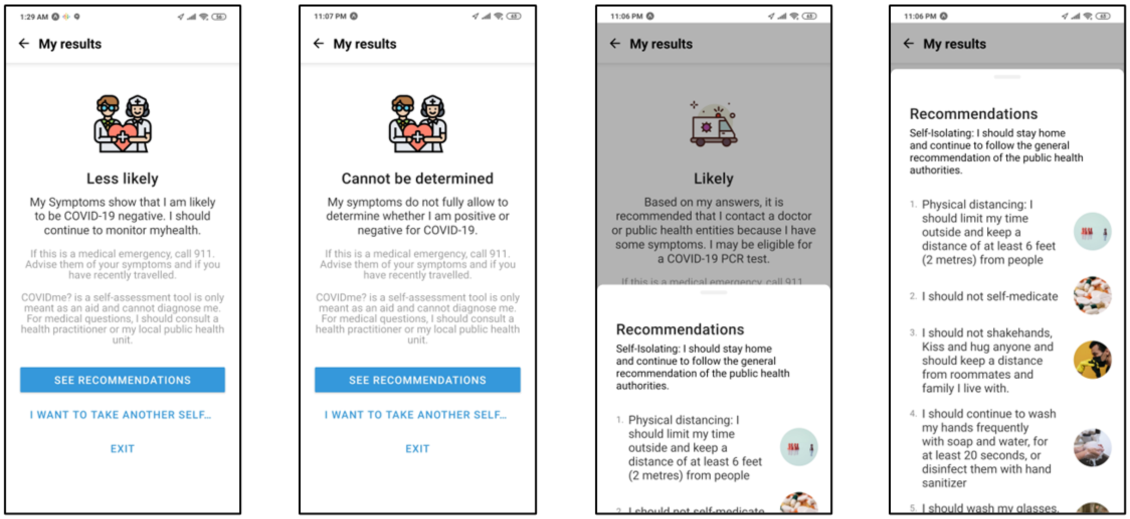
d. COVIDMe:
Survey summary

Figure 4-2. Screen camptures of the COVIDMe survey for collect symptoms as reported by the Physical Twin.

The implementation of COVIDMe has part of it deployed on the edge (i.e. a smartphone), and the Inference Engine was deployed on the cloud. Figure 4-2a shows the entry point for COVIDMe, asking for the PT preferred language. The PT will be more receptive to any recommendations from COVIDMe if these are provided in the PT's preferred language; this will enhance the response to the QoE-driven feedback loop.

Since COVIDMe was designed using the design ecosystem for PDTs for well-being, it also has a “Data Source” module responsible for data collection. This module is implemented on the edge (i.e., a smartphone); Figures 9b, 9c, and 9d comprise the “Data Source” component of COVIDMe as it is in these screens where data is being collected from the PT directly. Figure 4-2b is the starting point for data collection and prompts the PT to answer a built-in survey that presents a sequence of questions regarding the PT's state of health. The mobile application collects the data by prompting the user to report the presence or absence of symptoms via an in-app survey; Figure 4-2b shows the initial screen of the survey. Once the PT has started the survey, the “Data Source” module will present the PT with a sequence of questions regarding the current state of health of the PT; Figure 4-2c shows an example of these questions and there will be one screen for each one of the symptoms. Finally, when the PT has answered all the questions in the survey, the “Data Source” module will present a summary of the answers provided by the PT (Figure 4-2d); upon receiving confirmation that the answers are correct, these answers are stored locally.

The data collection module (i.e. data source) then transfers the data to the “Inference Engine” through the network. The “Inference Engine” will validate that the data is complete and validate it as an input for the model. The classifier deployed in the “Inference Engine” will use the collected data and produce a prediction. The prediction is required to provide feedback to the physical twin.



Less likely

My Symptoms show that I am likely to be COVID-19 negative. I should continue to monitor myhealth.

If this is a medical emergency, call 911. Advise them of your symptoms and if you have recently travelled.

COVIDme? is a self-assessment tool is only meant as an aid and cannot diagnose me. For medical questions, I should consult a health practitioner or my local public health unit.

Cannot be determined

My symptoms do not fully allow to determine whether I am positive or negative for COVID-19.

If this is a medical emergency, call 911. Advise them of your symptoms and if you have recently travelled.

COVIDme? is a self-assessment tool is only meant as an aid and cannot diagnose me. For medical questions, I should consult a health practitioner or my local public health unit.

Likely

Based on my answers, it is recommended that I contact a doctor or public health entities because I have some symptoms. I may be eligible for a COVID-19 PCR test.

Recommendations

Self-Isolating: I should stay home and continue to follow the general recommendation of the public health authorities.

1. Physical distancing: I should limit my time outside and keep a distance of at least 6 feet (2 metres) from people

Recommendations

Self-Isolating: I should stay home and continue to follow the general recommendation of the public health authorities.

1. Physical distancing: I should limit my time outside and keep a distance of at least 6 feet (2 metres) from people
2. I should not self-medicate
3. I should not shakehands, Kiss and hug anyone and should keep a distance from roommates and family I live with.
4. I should continue to wash my hands frequently with soap and water, for at least 20 seconds, or disinfect them with hand sanitizer
5. I should wash my glasses.

a. COVIDMe:
Not infected

b. COVIDMe:
Undetermined

c. COVIDMe:
Likely infected

d. COVIDMe:
Recommendations

Figure 4-3. Screen capture of COVIDMe QoE-driven feedback loop.

The feedback is presented in text and images in seconds after capturing the complete list of symptoms, as shown in Figure 4-3; Figure 4-3a shows the type of feedback a PT would receive in case the probability of infection is low, likewise, when the prediction is not decisive, the feedback for the PT would be like the one shown in figure 4-3b. In contrast, when a positive infection has been predicted, the PT will receive a series of recommended actions, as shown in Figures 10c and 10d. Additionally, the feedback module can show the results of previous assessments thanks to the COVIDMe Data Storage module. It is essential to mention that COVIDMe is not meant to replace a PCR test, which is the gold standard to diagnose COVID-

19; instead, among the objectives of this platform is to support the decision process for individuals and health jurisdictions when dealing with the spread infectious diseases like COVID-19 by creating a PDT for well-being. The feedback produced for the individual is one of three types: low probability, inconclusive and high probability. COVIDMe will then prompt the user with the corresponding recommendations. The next section shows a kite-level use case diagram for COVIDMe.

4.2.1. Use-Case Diagram

COVIDMe has two actors: the Physical Twin and the Digital Twin. Each of the use cases in Figure 4-4 is detailed next.

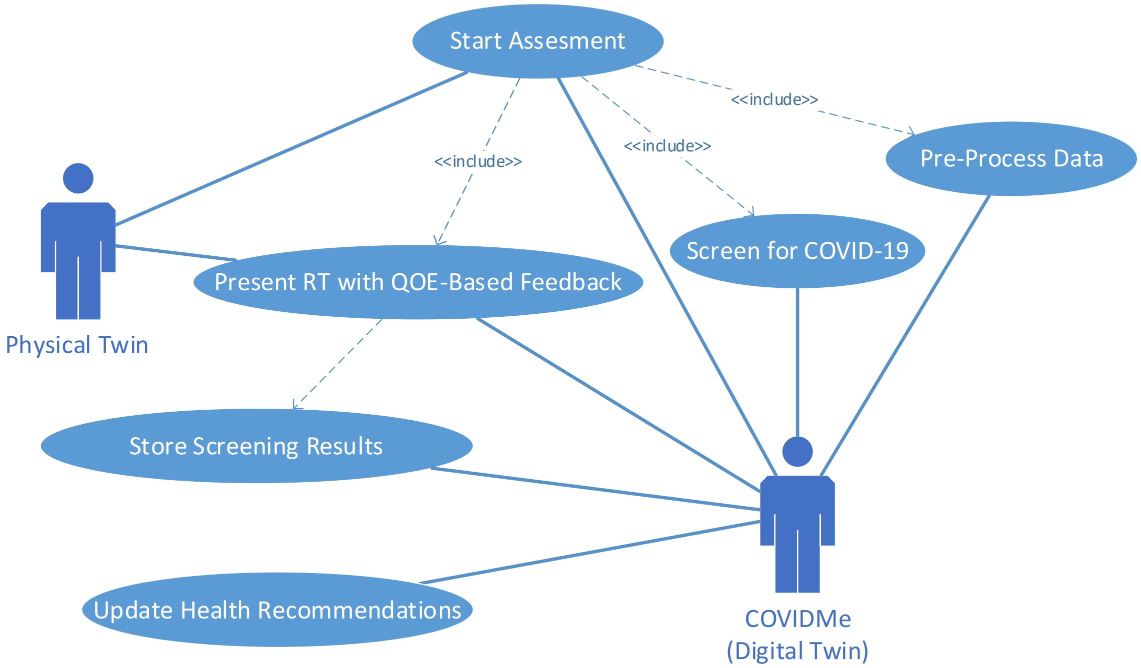


Figure 4-4. Kite level use case diagram for COVIDMe. Adapted from [85].

Start Assessment. The use case starts when COVIDMe prompts the PT to answer the survey of symptoms associated with COVID-19. The PT will immediately proceed to answer the survey and tap on the button "Send" when finished. Once the PT has responded to the study, COVIDMe (DT) will format the answers into a tuple of symptoms. At this point, the use case for "Pre-Process Data" will receive the tuple of symptoms to generate a feature array used in the "Screen for COVID-19" use case to create an assessment of the PT's health status. The evaluation is stored immediately in the "Store Screening" use case to generate appropriate feedback to the PT in the "Present PT with QOE-Based Feedback" parallel.

Pre-Process Data. The use case starts after COVIDMe receives the tuple of symptoms collected from the survey and encodes it into an array of features. It ends with the creation of various features.

Screen for COVID-19. The use case starts when COVIDMe feeds the feature array into the COVID-19 classifier to calculate the PT's probability of being infected with SARS-CoV-2 (COVID-19). The result is equivalent to screening for COVID-19, and it can be expressed in one of three states: the PT is likely sick of COVID-19, the PT is not likely sick of COVID-19 and Undetermined. The use case ends with the screening being created.

Store Screening Results. The use case starts when COVIDMe obtains a screening for COVID-19, which is sent to a repository in the cloud for further analysis.

Present PT with QOE-Based Feedback. The use case starts with COVIDMe presenting the screening process results as quality-of-experience-based feedback. COVIDMe will select proper health guidelines (recommendations) to advise the PT to follow the screening results.

Update Health Recommendations. The use case starts with COVIDMe harvesting guidelines/recommendations for the general population regarding COVID-19 and ends with COVIDMe updating its internal repository with the most recent guidelines.

4.2.2. Communication Diagram

A communication diagram is helpful to describe the dynamics between a system's different parts and components. Figure 4-5 shows nine lifelines (boxes), each representing an element of COVIDMe in a communication diagram. The whole platform is deployed in three different places. The PT smartphone runs the mobile app that holds the "Data Source" and "Multimodal Interaction" represented in the diagram by the "COVIDMe-App" lifeline. The "AI-Inference

Engine" module is deployed in a cloud computing service. It can be any if it supports the deployment of Docker© containers. The lifeline corresponding to this bit of COVIDMe is "COVIDMe-Microservice." The data repository is also deployed in the cloud.

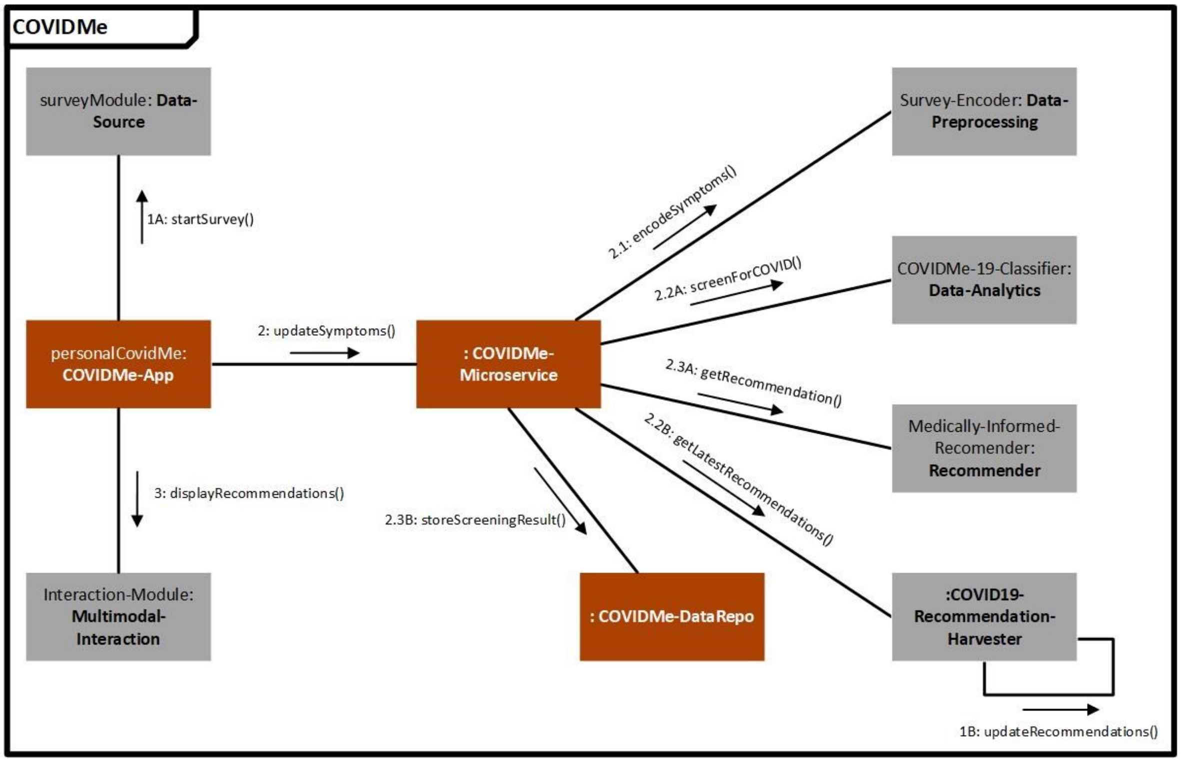


Figure 4-5. Low-granularity Communication diagram for COVIDMe. Recovered from [85].

Lifelines interact with each other, and in most cases, these interactions happen in a specific order. The diagram shows the interactions in the form of messages. One way to read it is by following the sequence of each message. Table 4-1 describes the interactions triggered by each message.

Table 4-1. Detailed descriptions of the messages for the low-granularity COVIDMe communication diagram.

Numeral	Message	Description
1A	startSurvey()	The COVIDMe-App will trigger the Data-Source module to start the survey to collect the symptoms from the PT.
1B	updateRecommendations()	The Recommendation Harvester will look for the most up-to-date recommendations for self-managing COVID-19 in different situations and at official sources.
2	updateSymptoms()	The COVIDMe-App will send the COVIDMe-Microservice a new tuple with the most recent symptoms reported by the PT.
2.1	encodeSymptoms()	The COVIDMe-Microservice will preprocess the tuple of symptoms to transform it into an array of features.
2.2A	screenForCOVID()	The COVIDMe-Microservice will send the array of features to the COVIDMe-19-Classifier, which will classify the features and calculate the PT's probability of having COVID-19.
2.2B	getLatestRecommendations()	The COVIDMe-Microservice will recover the latest recommendations and guidelines to manage COVID-19 from the Recommendation-Harvester.
2.3A	getRecommendation()	Based on the prediction from 2.2A, the COVIDMe-Microservice will get the corresponding recommendation to display to the PT.
2.3B	storeScreeningResult()	The COVIDMe-Microservice will send the prediction from 2.2A to the COVIDMe-DataRepo for storage.
3	displayRecommendations()	The COVIDMe-App will send the corresponding recommendations from 2.3A to be displayed by the Multimodal-Interaction lifeline as a form of human-readable feedback for the PT.

The next section will discuss the contributions and open challenges for a DT for health in COVID-19.

4.3. Machine learning modelling

This section details the training and evaluation of 15 different classifiers with learning algorithms implemented in Scikit-learn [86], such as decision trees, neural networks, and support vector machines. The algorithm used to train the support vector machine classifier is the C-Support Vector Classifier, which implements libsvm for Scikit-learn. In addition to the three algorithms, two types of ensembles: random forest (RF) and a voting classifier with 2, 3, and 4 votes using the three algorithms and the random forest classifier as votes. The experiments ran on a computer with a processor intel core-i7 @ 2.6 GHz, 16 GB of RAM, and a GPU with 640 CUDA cores. As for the datasets, the public health authorities from Sonora and Tlaxcala in Mexico compiled two different datasets from medical records collected from patients suspected of being infected with SARS-CoV-2. Each state compiled one dataset from its patients separately using the same variables. The two datasets were merged and then processed to obtain a single dataset with 166924 complete data samples collected from 83462 COVID-19-positive and 83462 COVID-19-negative patients.

The ML modelling process required exploring various learning algorithms to discover the one that renders the best results. Decision Trees, Neural Networks, and Support Vector Machines are explained next to provide insight into how these learning algorithms work.

4.3.1. Decision tree

Breiman et al. [87] introduced the decision tree learning algorithm in a study titled Classification and Regression Trees (CART); the learning algorithm has the same name. CART will use the independent variable to divide the dataset into subsets of instances, each subset corresponding to a class. The criteria for separating the dataset is to use the feature in the dataset that better splits the same into homogeneous subsets. Each subset becomes a branch in the tree, and CART will grow more branches from it recursively until it reaches the maximum allowed depth or produces an entirely homogeneous subset. For a binary classification problem, CART grows a binary tree. This chapter addresses the problem of SARS-CoV-2 detection from self-reported symptoms, which is a binary class classification problem, infected or not infected. If applied to this problem, CART would find the symptom that produces the most homogeneous split in the dataset to produce two branches. Let us imagine that the symptom that does the best split is fever, and those examples who have a fever are assumed to be infected, while those who do not have a fever are considered infection-free. The CART algorithm will recursively grow more branches from that point, and for the next level, it will take other symptoms to do more splits. The key is finding the feature that produces the most homogeneous separation; this is assessed by measuring the impurity of the division. For this purpose, the gini index or gini impurity function is used.

The Gini index is a measurement that indicates the probability of a particular sample being misclassified with a given split. The CART algorithm uses the gini index to decide the best split to grow the decision tree branches. For example, suppose there is a dataset with three independent variables or features and two possible classes to be assigned to a sample. In that case, CART will measure the impurity of a split made with each of the three features and select the one that produces the less impure division. The following formula corresponds to the gini index.

$$Gini\ Index = 1 - \sum_{i=1}^{n} (p_i)^2$$

In the previous formula, p is the probability of a sample being assigned to a particular class i. The learning algorithm will measure the gini index and select the feature that creates the split with the lower impurity or misclassification probability, which will be calculated with the gini index for each branch and each tree level.

4.3.2. Neural network

Neurons are the building blocks of neural networks. A neuron has input values added and fed to an activation function to generate an output. Each input has a weight that amplifies or reduces an input value. Depending on the application, a neuron may adopt different activation functions: threshold function, rectifier, linear unit, hyperbolic tangent, and sigmoid function. The mathematical definition for a neuron would be as follows.

$$y = \varphi\left(\sum_{i=1}^{m} w_i x_i\right)$$

Where y is the output value of a given neuron and is calculated by evaluating the activation function φ on the sum of all weighted inputs $w \cdot x$, the neural network is an array of neurons organized in layers composed of any number of networks. In some cases, when a neural network has a large number of layers, it then becomes a deep learning problem. The outputs of a layer are inputs for the next layer. The sigmoid function is usually preferred as an activation function as it has a codomain of real numbers $0<y,$ which allows us to play with the classification threshold and obtain optimal performance. The sigmoid function in the Scikit-learn library is implemented as follows.

$$\frac{1}{1+e^{-x}}$$

The weights in a neural network are adjusted to amplify or reduce the effect a neuron has in the output of the final layer, which is accomplished utilizing a solver or weight optimization algorithm, the most popular being adam [88].

4.3.3. Support vector machines

Support vector machines (SVM) is a simple yet powerful machine learning approach. The intuition of SVM is to map each sample in the dataset to a multidimensional space by assigning each feature to a dimension. Thus, the independent variables are also known as dimensions for the SVM. The next step is to define a boundary delimiting the region in the space corresponding to each class in the dataset. This boundary has the form of a hyperplane in the multidimensional space. Mathematically, a hyperplane is a line, a polynomial, or a radial function. It is used to decide if a new sample belongs to a class or not, depending on where in the space it is mapped.

4.3.4. Voting ensembles

A voting classifier is an ensemble of two or more estimators (classifiers). Each estimator will issue a prediction for all classes. In a binary problem like the one described in this study, each estimator will also predict the probability for a positive and a negative class. If a hard vote is configured for the voting classifier, the final prediction will be based on a majority vote. However, each estimator is assigned a weight, so it may be the case that each estimator contributes differently to the final prediction. When the voting classifier is set to a soft vote, the final prediction is the result of applying an arg max function to the probabilities of each estimator.

4.3.5. Datasets

A successful classifier can generalize so that it can correctly classify new samples. The dataset's quality significantly affects the model's capacity to generalize. For this study, the data came directly from clinics and hospitals in the public health system in Sonora and Tlaxcala, and health professionals collected it. This guarantees that the collected data is accurate and that the collection process follows medical guidelines for health practitioners in clinics and hospitals in Mexico. Both datasets were merged into a single, homogeneous dataset since both were collected following the same medical guidelines put in place by health authorities throughout the country (Mexico). The public health authorities in both states collect several variables for each sample. These variables include 20 symptoms, comorbidities, signs like sudden onset of symptoms and laboratory test results for RT-PCR for SARS-CoV-2. Predicting a positive infection based on self-reported symptoms is paramount to removing the need for the patient to visit a laboratory to provide samples just to get screened, as self-reported symptoms can be collected remotely. Symptoms, data reporting the sudden onset of symptoms and known contact with confirmed cases were selected as features for training and validating the models.

The combined dataset has 215786 medical records that account for the same number of patients (i.e., one per patient). In addition to patients infected with SARS-CoV-2 and patients that tested negative for any sort of viral infection, the 215786 samples also include confirmed patients infected with influenza or coronavirus. The dataset was processed to keep an equal number of samples indicating SARS-CoV-2 and negative to any infection. The objective of the ML model is to detect OCVID-19 infections based on self-reported symptoms, risk factors and signs. The dataset had 22 features that fit this criterion: 20 are symptoms, and two more features report any known contacts with COVID-19-positive confirmed cases and the sudden onset of the symptoms. After identifying these features as independent variables, the dataset was processed to filter-out all incomplete records with empty or unknown values in any of the twenty-two variables. The dependent or target variable is the result of the RT-PCR. After undersampling the majority class (negative cases), the resulting dataset had 166924 data samples from 83462 COVID-19-positive and 83462 COVID-19-negative patients. This process results in a dataset with enough records for training and validation.

4.3.6. Performance metrics

Before talking about performance metrics, true-positive, false-negative, true-negative and false-positive concepts are necessary. In a binary classification problem such as the one addressed in this study, True-Positive (TP) is the number of true instances that the model correctly classifies. False-Negative (FN) is the number of true samples the model misclassifies. True-Negative (TN) is the number of negative samples the model correctly classifies. Finally, false-positive (FP) is a negative instance that the model misclassifies.

Sensitivity is the rate of true positive instances correctly classified compared to the total number of true instances. The formula to calculate sensitivity is as follows,

$$Sensitivity = \frac{tp}{tp + fn}$$

Specificity is the rate of true-negative instances that are correctly classified by the model when compared with the total number of negative samples as follows,

$$Specificity = \frac{tn}{tn + fp}$$

Precision is the rate of true-positive samples when compared with the number of instances classified as positive. The formula is as follows,

$$Precision = \frac{tp}{tp + fp}$$

In many classification problems, knowing how our model's sensitivity compares to precision is important. F1 score can have a value between 0 and 1. A perfect F1 score means that all samples classified as positive are, in fact, positive (no false positives) and that all positive samples in the dataset are correctly classified by the model (no false negatives). Ideally, this is achieved when the model is properly trained. The formula goes as follows,

$$F1 = \frac{2 \times (sensitivity \times precision)}{(sensitivity + precision)}$$

Accuracy is the rate of correctly classified samples compared to the complete dataset. Accuracy by itself in gauging the performance of a model can be misleading as the model could have a high specificity and a low sensitivity, resulting from a poorly trained model. The formula is as follows,

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Sometimes, depending on which is more relevant to the classification task: the true-positive rate (sensitivity) or false-positive rate (specificity), the threshold at which a model classifies a sample as positive or negative may be adjusted. That is precisely how the receiver operating characteristic (ROC) curve is plotted. The ROC curve is a good metric to assess the quality of a model and decide on adjusting the classification thresholds so that the model renders the best scores for a particular classification problem and circumstance. The area under the ROC curve (ROC_AUC) is used to assess the performance of all the models reported in this study. Ideally, a properly trained model would get close to 1. A classifier with the highest sensitivity and the lowest false-positive rate is the best.

4.3.7. Feature importance analysis

Feature importance is an impurity-based measurement representing how important a feature is in predicting the target variable. The gini index is used to calculate the importance of all the features in the dataset, which is why this technique is also known as gini importance. This measurement, also known as gini importance, is calculated by fitting the dataset to a decision tree classifier. Ensembles based on decision trees can calculate the importance of features as well. Table 4-2 shows the results of a feature importance analysis made on the dataset.

Table 4-2. Results of the feature importance analysis conducted on 22 independent variables representing self-reported symptoms in the COVIDMe dataset along with baseline ROC_AUC scores resulting from fitting the dataset to a non-tunned random forest classifier by progressively including features; the items are sorted according to ranking of importance.

Ranking	Importance	ROC_AUC	Feature name	Behaviour (ROC_AUC)
1	0.0829	0.6328	FEVER	-----
2	0.0616	0.6510	COUGH	+
3	0.0598	0.6396	SUDDEN_ONSET_OF_SYMPTOMS	-
4	0.0579	0.6841	KNOWN_COVID_CONTACT	+
5	0.0573	0.6947	ODYNOPHAGIA	+
6	0.0558	0.6976	RHINORRHEA	+
7	0.0535	0.6713	IRRITABILITY	-
8	0.0519	0.6826	CHEST_PAIN	+
9	0.0517	0.6664	CEPHALEA	-
10	0.0516	0.6552	CHILLS	-
11	0.0501	0.6687	DIARRHEA	+
12	0.0478	0.6764	FATIGUE_LOSS_APPETITE	+
13	0.0443	0.6905	DYSPNEA	+
14	0.0435	0.7005	MYALGIA	+
15	0.0425	0.6966	ARTHRALGIA	-
16	0.0403	0.6945	CONJUNCTIVITIS	-
17	0.0370	0.7010	ABDOMINAL_PAIN	+
18	0.0293	0.7122	POLYPNEA	+
19	0.0275	0.7064	VOMIT	-
20	0.0235	0.7098	ANOSMIA	+
21	0.0185	0.6864	DYSGEUSIA	-
22	0.0115	0.6985	CYANOSIS	+

The Centers for Disease Control and Prevention (CDC) report as symptoms of COVID-19 fever, chills, cough, shortness of breath, headache, fatigue, muscle or body aches, loss of smell or taste, sore throat, congestion, diarrhea and vomiting, although in no particular order of importance for the diagnose [89]. The Table 4-2 presents those symptoms on column “Feature name.” One approach to feature selection is to select the features with the highest gini importance scores and leave out the rest. However, adopting that approach, would leave out some of the symptoms associated with COVID-19 by the CDC. Thus, an alternative way to select the features is required. Ideally, a model will predict 100% of true-positive cases with 0% of false-positive cases. However, that is not realistic, so the next best thing is the model that has the highest possible true-positive rate and the lowest possible false-positive rate. This equates to find the model that best detect as many real cases of infected patients with the lowest possible rate of false alarms. The area under the curve for the receiver operating characteristic (ROC_AUC) is the ideal metric for these cases. A baseline performance was set using a random forest classifier by gradually including the features according to their importance ranking from Table 4-2. The process to set the baseline performance was to increasingly include the features from the twenty-two most important features in Table 4-2 to form a training dataset. First fever alone was used to train and validate a classifier; the ROC_AUC for that model was 0.6328. After, fever and cough were used to train a new model, the top two most important features from Table 4-2; which had an ROC_AUC value of 0.6510. Each time, a new symptom was included to train a new model, selected from the most important to the least important in Table 4-2. The process was repeated until all 22 features were included and the 22 models were validated. Each model was validated using 5-fold cross-validation. As shown in Table 4-2, when some of the symptoms were included to train a model, the ROC_AUC value drop. Intuitively, the best subset of features from Table 4-2 to use to build a training dataset, would be those that increased the value of the ROC_AUC. Another strategy to find the best subset of features to use to build a training dataset would be to try every possible combination. This implies to try all possible combinations of one feature, then all possible combinations of two features, then combinations of three features and so on to all possible combinations of 21 features out of the total 22. This was not possible due to limitations in computing resources. Due the computation limitations, the subset of features built with the ones that increase the ROC_AUC value is a good candidate to create a training dataset; also the full set of twenty-two features was used to train and evaluate classifiers. The results are shown in this chapter. The list of the features that contribute to increase the ROC_AUC value include those ranked in these positions (Table 4-2): 1, 2, 4, 5, 6, 8, 11, 12, 13, 14, 17, 18, 20, and 22.

4.3.8. Baseline scores

Both dataset versions have the same 1772 samples, each with 886 SARS-CoV-2 positive and 886 negative patients. Four learning algorithms were selected to train classifiers with the training datasets: decision tree, neural networks, support vector machine, and random forest. A voting classifier can outperform the classifiers that are part of it, so all possible combinations of two, three and four votes were explored. Table 4-3 shows the baseline scores for the resulting 15 different classifiers. The baseline hyperparameter configuration for the classifiers is the default configuration, as defined in Scikit-learn [21]. The voting classifier built with the baseline NN and SVM reports a ROC_AUC score of 0.716.

Table 4-3. Baseline scores resulting from fitting the COVIDMe dataset to non-tunned classifiers; results from fitting the classifiers with 14 and 22 features selected based on importance.

CLASSIFIER NAME	FEATURES	SENSITIVITY	SPECIFICITY	PRECISION	F1	ACCURACY	MEAN_AUC	STD_AUC
DECISION TREE	14	0.536	0.699	0.642	0.605	0.617	0.622	0.064
NEURAL NETWORK	14	0.699	0.631	0.655	0.661	0.665	0.712	0.040
RANDOM FOREST	14	0.636	0.674	0.661	0.653	0.655	0.680	0.054
SUPPORT VECTOR MACHINE	14	0.734	0.618	0.659	0.670	0.676	0.712	0.050
VOTING DT & NN	14	0.611	0.661	0.644	0.632	0.636	0.672	0.052
VOTING DT & RF	14	0.590	0.669	0.641	0.624	0.630	0.655	0.058
VOTING DT & SVM	14	0.620	0.649	0.639	0.632	0.634	0.670	0.052
VOTING DT, NN & RF	14	0.607	0.670	0.648	0.635	0.639	0.678	0.055
VOTING DT, NN & SVM	14	0.631	0.649	0.643	0.637	0.640	0.692	0.053
VOTING DT, NN, SVM & RF	14	0.636	0.656	0.647	0.644	0.646	0.691	0.052
VOTING DT, SVM & RF	14	0.612	0.662	0.644	0.633	0.637	0.680	0.054
VOTING NN & RF	14	0.673	0.668	0.670	0.669	0.670	0.702	0.048
VOTING NN & SVM	14	0.703	0.629	0.656	0.662	0.666	0.716	0.040
VOTING NN, SVM & RF	14	0.684	0.641	0.656	0.660	0.662	0.707	0.050
VOTING SVM & RF	14	0.672	0.652	0.660	0.661	0.662	0.698	0.050
DECISION TREE	22	0.556	0.668	0.629	0.605	0.612	0.606	0.039
NEURAL NETWORK	22	0.649	0.643	0.647	0.643	0.646	0.690	0.046
RANDOM FOREST	22	0.666	0.631	0.645	0.645	0.648	0.703	0.043
SUPPORT VECTOR MACHINE	22	0.743	0.609	0.656	0.668	0.676	0.718	0.048
VOTING DT & NN	22	0.585	0.642	0.622	0.610	0.613	0.679	0.050
VOTING DT & RF	22	0.585	0.638	0.620	0.607	0.611	0.681	0.045
VOTING DT & SVM	22	0.595	0.635	0.622	0.612	0.615	0.688	0.049
VOTING DT, NN & RF	22	0.619	0.661	0.647	0.638	0.640	0.696	0.048
VOTING DT, NN & SVM	22	0.625	0.644	0.639	0.633	0.635	0.700	0.046
VOTING DT, NN, SVM & RF	22	0.658	0.649	0.653	0.651	0.654	0.708	0.046
VOTING DT, SVM & RF	22	0.601	0.643	0.630	0.618	0.622	0.696	0.043
VOTING NN & RF	22	0.666	0.616	0.635	0.637	0.641	0.708	0.050
VOTING NN & SVM	22	0.692	0.616	0.644	0.649	0.654	0.713	0.049
VOTING NN, SVM & RF	22	0.699	0.616	0.647	0.652	0.657	0.720	0.046
VOTING SVM & RF	22	0.702	0.623	0.652	0.657	0.662	0.725	0.043

4.3.9. Hyperparameter tuning

This section explains our approach to hyperparameter tuning for the 15 classifiers. After tuning each model from the baseline configuration, each model was evaluates using 10-fold cross-validation against both datasets: 14 and 22 features. The results are shown in Table 4-4.

Table 4-4. Scores after fine-tunning classifiers fitted with the COVIDMe dataset.

CLASSIFIER NAME	NUMBER OF FEATURES	SENSITIVITY	SPECIFICITY	PRECISION	F1	ACCURACY	ROC_AUC	STD_AUC
DECISION TREE	14	0.620	0.667	0.651	0.639	0.643	0.686	0.038
NEURAL NETWORK	14	0.726	0.619	0.658	0.665	0.673	0.721	0.044
RANDOM FOREST	14	0.738	0.630	0.667	0.678	0.684	0.723	0.046
SUPPORT VECTOR MACHINE	14	0.710	0.639	0.664	0.670	0.674	0.711	0.047
VOTING DT & NN	14	0.712	0.630	0.659	0.667	0.671	0.710	0.037
VOTING DT & RF	14	0.716	0.642	0.669	0.675	0.679	0.719	0.046
VOTING DT & SVM	14	0.707	0.633	0.660	0.665	0.670	0.708	0.047
VOTING DT, NN & RF	14	0.715	0.628	0.661	0.666	0.672	0.721	0.046
VOTING DT, NN & SVM	14	0.718	0.631	0.661	0.669	0.674	0.722	0.045
VOTING DT, NN, SVM & RF	14	0.717	0.634	0.664	0.671	0.675	0.722	0.041
VOTING DT, SVM & RF	14	0.726	0.631	0.664	0.673	0.678	0.719	0.047
VOTING NN & RF	14	0.733	0.625	0.662	0.674	0.679	0.724	0.045
VOTING NN & SVM	14	0.716	0.634	0.663	0.670	0.675	0.718	0.038
VOTING NN, SVM & RF	14	0.722	0.631	0.663	0.671	0.677	0.721	0.047
VOTING SVM & RF	14	0.733	0.633	0.668	0.677	0.683	0.721	0.048
DECISION TREE	22	0.747	0.552	0.631	0.622	0.650	0.693	0.040
NEURAL NETWORK	22	0.682	0.638	0.654	0.655	0.660	0.718	0.051
RANDOM FOREST	22	0.729	0.607	0.652	0.660	0.668	0.730	0.047
SUPPORT VECTOR MACHINE	22	0.788	0.569	0.647	0.659	0.678	0.719	0.048
VOTING DT & NN	22	0.708	0.626	0.656	0.662	0.667	0.726	0.042
VOTING DT & RF	22	0.742	0.601	0.652	0.661	0.672	0.730	0.044
VOTING DT & SVM	22	0.763	0.591	0.652	0.665	0.677	0.726	0.040
VOTING DT, NN & RF	22	0.714	0.623	0.657	0.663	0.669	0.722	0.050
VOTING DT, NN & SVM	22	0.712	0.618	0.653	0.658	0.665	0.721	0.041
VOTING DT, NN, SVM & RF	22	0.727	0.619	0.658	0.665	0.673	0.729	0.045
VOTING DT, SVM & RF	22	0.739	0.608	0.655	0.665	0.674	0.728	0.046
VOTING NN & RF	22	0.713	0.618	0.653	0.660	0.666	0.728	0.049
VOTING NN & SVM	22	0.721	0.618	0.656	0.664	0.670	0.726	0.050

VOTING NN, SVM & RF	22	0.726	0.622	0.660	0.667	0.674	0.730	0.048
VOTING SVM & RF	22	0.742	0.608	0.656	0.665	0.675	0.729	0.048

4.3.10. Hyperparameter optimization

There are different approaches to searching for the optimal hyperparameter configuration to train a machine learning model, namely Particle Swarm Optimization (PSO), Bayesian Optimization (BO), Genetic Algorithms (GA), Grid Search (GS) and Random Search (RS). The grid search method defines a range of all possible values for a set of hyperparameters. It tests all possible permutations to find the best configuration to train a machine learning model. If followed thoroughly, this approach guarantees to find the optimal configuration of hyperparameters for a particular learning problem. All possible permutations of different values for a defined set of hyperparameters are known as search space. The downside of grid search is that training a model with the entire search space of hyperparameters is computationally expensive and, in most cases, impractical; its implementation's simplicity is its primary advantage. An alternative to it is a random search. This method selects a random distribution of the entire search space of hyperparameters to test as candidates. This approach does not guarantee to find the optimal solution for a particular problem, but a close-to-optimal solution [90]. The optimal or at least close-to-optimal hyperparameter configuration for the decision tree, neural network, support vector machine and random forest classifiers was found using random search. However grid search was used for the ensembles.

We implemented a random search to tune the hyperparameters of a decision tree, neural network, support vector machine and random forest classifiers. After the optimization of the models made with random forest, neural network, support vector machine, and decision tree, ensembles were compiled using the optimized classifiers. A grid search strategy was adopted to optimize the ensembles. Table 4-5 defines the search spaces for all classifiers.

Table 4-5. Search space used for fine-tunning the COVIDMe classifiers applying random search.

Decision tree	
Hyperparameters	Range of values
criterion	['gini', 'entropy']
splitter	['best', 'random']
max_depth	[50 values equally distributed between 2 and 103 (inclusive)]
min_samples_split	[2, 3, 4, 5, 6, 7, 8, 9, 10]
min_samples_leaf	[2, 3, 4, 5, 6, 7, 8, 9, 10]
max_features	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
ccp_alpha	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Search space:	1782000
Random forest	
n_estimators	[500 equally distributed values between 2 and 2003 (inclusive)]
criterion	['gini', 'entropy']
max_depth	[50 equally distributed values between 2 and 103 (inclusive)]
min_samples_split	[2, 3, 4, 5, 6, 7, 8, 9, 10]
min_samples_leaf	[2, 3, 4, 5, 6, 7, 8, 9, 10]
# max_features	['auto', 'sqrt', 'log2', None]
max_features	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
bootstrap	[True, False]
ccp_alpha	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Search space:	3564000000
Support vector machine	
C	[40 equally distributed values between 0 and 2 (inclusive)]
kernel	['linear', 'poly', 'rbf', 'sigmoid']
degree	[1, 2, 3, 4, 5, 6]
gamma	[100 equally distributed values between 0.000000001 and 0.999 (inclusive)]
coef0	[200 equally distributed values between 0 and 1 (inclusive)]
shrinking	[True, False]
Search space:	38400000
Neural network	

hidden_layer_sizes	[tuples of 1, 2 and 3 layers each with 15, 50, 80, 100 and 200 neurons]
activation	['identity', 'logistic', 'tanh', 'relu']
solver	['lbfgs', 'sgd', 'adam']
alpha	[500 equally distributed values between 0.00001 and 0.001 (inclusive)]
learning_rate	['constant', 'invscaling', 'adaptive']
learning_rate_init	[20 equally distributed values between 0.00001 and 0.01 (inclusive)]
power_t	[10 equally distributed values between 0.1 and 1.0 (inclusive)]
shuffle	[True, False]
Search space:	1116000000
Voting classifier (two votes)	
estimators	[6 different combinations of two estimators using the best DT, NN, SVM and RF]
weights	[16 different tuples of two integer values between 1 and 4 (inclusive)]
Search space:	96
Voting classifier (three votes)	
estimators	[4 different combinations of two estimators using the best DT, NN, SVM and RF]
weights	[64 different tuples of three integer values between 1 and 4 (inclusive)]
Search space:	256
Voting classifier (four votes)	
estimators	[1 tuple made of all four estimators: DT, NN, SVM and RF]
weights	[256 different tuples of four integer values between 1 and 4 (inclusive)]
Search space:	256

4.3.11. Modelling results

As shown in Table 4-3, the baseline ROC_AUC score for the decision tree classifier is 0.622 when trained with the 14 features dataset and 0.606 for the dataset with 22 features. After hyperparameter optimization with random search, the ROC_AUC score increased to 0.686 for the 14 features dataset and 0.693 for the 22 features dataset, as shown in Table 4-4. The random search was performed on 0.3% of the search space, as shown in Table 4-5, which means that

5346 different configurations of hyperparameters were tested. The baseline ROC_AUC for the random forest classifier is 0.68 when trained with the 14 features dataset and 0.703 with the 22 features dataset, as shown in Table 4-1. Table 4-5 shows the search space for hyperparameter tuning; the ROC_AUC for the 14 features dataset was 0.723 and 0.730 for the 22 feature dataset, as shown in Table 4-4. The neural network classifier had a baseline ROC_AUC of 0.712 and 0.690 for the 14 and 22 feature datasets, respectively (table 4-1). After hyperparameter optimization with random search, the scores for ROC_AUC were 0.721 for the 14 features dataset, while the score for the 22 features dataset was 0.718, as reported in Table 4-4. Support vector machines trained with the 14 and 22 features had baseline scores of 0.712 and 0.718, as reported in Table 4-3. After hyperparameter tunning, the scores were 0.711 for the 14 features dataset and 0.719 for the 22 features dataset, as shown in Table 4-3.

We tuned and validated 11 voting classifiers assembled from all possible combinations of four tunned classifiers: decision tree, neural network, support vector machine, and random forest. All voting classifiers were set to a soft vote to enable a probability prediction necessary to obtain a ROC_AUC score. A hard vote approach forces the classifier to make a prediction based on a majority vote. In contrast, a soft vote approach makes the voting classifier predict based on an arg max of the predicted probabilities from each classifier. For a voting classifier, the two most crucial parameters to configure are the estimators or classifiers and the weights assigned to each estimator. The ensembles were complied using the top best hyperparameter configurations for each of the classifiers, namely, decision tree, neural network, support vector machine and random forest; combinations of 2, 3 and 4 estimators were tested. These combinations are passed as an array of estimators to the ensemble. Details of the search space are shown in Table 4-5; however, each estimator takes a weight of 1, 2, 3 or 4 in all voting classifiers. Again, a grid search approach was adopted to optimize the hyperparameters with both versions of the datasets (i.e., 14 and 22 features) on each one of the 15 voting classifiers. Once the ensembles were tunned, a 10-fold cross-validation against the 14 and 22 feature datasets was conducted; the validation results are summarized in Table 4-4. The voting classifier built with the tunned decision tree and random forest classifiers reports the highest ROC_AUC score: 0.73 against the 22 feature dataset, as reported in Table 4-4.

Table 4-6. Scoring of the best classifier for COVID-19 prediction.

Metric	**22 features**
accuracy	0.68112423
classifier	voting DT & RF
f1	0.670486031
roc_auc	0.728663965
precision	0.660438774
sensitivity	0.752872829
specificity	0.609307967
std_auc	0.042059802

General practitioners must decide whether a patient has a flu-like disease or a severe acute respiratory infection. This prediction is exclusively based on symptoms and recorded in the attribute PROBABLE_DIAGNOSE. Plotting the receiving operating characteristic curve or calculating the area under the curve is not feasible for predictions made by general practitioners.

As shown on Table 4-4, the support vector machine on the 22 features has the highest sensitivity of 0.788. Still, at the same time, 43.1% of negative patients respectively are predicted to be infected with SARS-CoV-2. The false positive rate creates a false alarm for many non-infected patients. This problem requires that the classifiers yield the lowest number of false alarms possible and detects the highest number of infected people. There is a trade-off between these two metrics: the proportion of correct detections in relation to the false alarms. The f1 metric is precisely used to measure this trade-off. Table 4-6 summarizes the scores for the model that presented the highest ROC_AUC and f1 values in the ensemble of DT and RF with 22 features. The mean ROC_AUC score was 0.728, with a standard deviation of 0.042, a mean sensitivity of

0.752, and a mean specificity of 0.609. With these results COVIDMe may not compare directly to the gold standard, a RT-PCR test however, the results from the model evaluation proved that it can be used for initial screening as it had better results that the initial assessment done by physicians before the test.

After a validation on the model, COVIDMe was validated in a trial run using data samples and 20 people.

4.4. COVIDMe integration test

The primary objective of this integration test was to evaluate the overall functionality and predictive accuracy of the COVIDMe mobile application. The test was structured as a comprehensive integration evaluation involving real user interactions with the mobile application under controlled conditions. The aim was to verify the application's performance in terms of its operational integration and the consistency of its machine learning classifier in predicting COVID-19 infections accurately.

Twenty individuals participated in the trial, each tasked with simulating the interaction with the application by inputting symptom data from a predefined set of real patient records. The materials for this integrations test are:

- The Mobile Application (COVIDMe): Incorporates a user interface for symptom input and displays predictions of infection probability.
- Patient Data: A dataset of 200 real patient records, including self-reported symptoms and confirmed COVID-19 status via PCR tests, was used. These records were not included in the dataset used to train the application's underlying machine learning classifier to ensure the validity of the test by using unseen data.

Participants were briefed on the test procedures and the functionality of the COVIDMe app. Each participant was assigned 10 unique patient records from the dataset of 200 records.

The participants entered the symptoms from their assigned patient records into the COVIDMe app. This step was repeated for each record, with participants recording the predicted infection probability output by the app for each entry.

For each prediction made by the app, the participants recorded the output. None of the participants knew the real PCR result for each of the cases they were given to blind the participants to the real result form the PCR test. If the participants obtained a prediction from COVIDMe for each one of the 10 real cases, then the integration was considered successful.

The performance of the COVIDMe application was evaluated based on:

- Sensitivity: The ability of the app to correctly identify positive cases of COVID-19 as positive.
- Specificity: The ability of the app to correctly identify negative cases of COVID-19 as negative.
- Accuracy: The overall accuracy of the app in predicting the correct infection status.

These metrics were calculated by comparing the app's predictions with the actual PCR test results from the patient records. The test confirmed that the COVIDMe app maintained consistent sensitivity, specificity, and accuracy in predicting COVID-19 status across all trials with the real patient data. This indicated that the machine learning classifier functioned reliably within the integrated mobile application environment.

The integration test demonstrated that the COVIDMe application, as a Personal Digital Twin for COVID-19, operates effectively in a real-world scenario, accurately processing user-inputted symptoms and generating reliable predictions of COVID-19 infection. The application proved to be a robust tool, integrating user interface, data processing, and machine learning components seamlessly to support health decision-making processes.

This classifier is not intended to replace an RT-PCR test, nor outperform the primary care practitioners' assessment. The purpose of implementing COVIDMe is:

- To explore the problem of atomate the detection of viral infections with a PDT for well-being.
- As health services worldwide continue to face challenges with fluctuating demands from various infectious diseases, governments and health authorities are increasingly exploring innovative, non-conventional approaches to reduce infection rates. These strategies encourage individuals to take proactive steps in managing their health, thereby reducing reliance on medical services. This chapter introduces a novel approach through COVIDMe, originally developed during the COVID. However this can be adapted as a broader machine learning tool capable of assessing the probability of infection from a range of infectious diseases. By providing immediate and personalized health assessments, COVIDMe empowers individuals to take preventive actions, significantly helping to curb the spread of infectious diseases and enhance public health resilience.

Experienced primary care practitioners (i.e., physicians) often reach accurate diagnoses based on signs, symptoms, and past experiences. Something similar can be said for machine learning algorithms; the number and quality of the observations play a paramount role in the performance of the resulting model, especially when the model is predicting the infection of a disease like COVID-19. COVIDMe contributes a machine-learning model that detects infected COVID-19 patients with an average sensitivity of 0.752. Another contribution is the integration of revision and aggregation of two COVID-19 datasets into one, which is available for other data scientists to explore and contribute to this problem.

One of the limitations of this study is that COVIDMe is at a proof-of-concept stage and has been pilot-tested only with a handful of users. Using self-reported symptoms to detect COVID-19 mainly because it is easy to have participants send data without needing to leave their homes, thus keeping participants safe. However, we need to explore more data types in the future and find which approach for a classifier is best.

COVIDMe needs standards; the self-reported symptoms, for instance, are not organized in any particular way. That means the developer needs to be sure of the order in which the classifier expects that input. By standardizing the data in the platform, we may ensure compatibility between different components, even if these are updated or swapped for other implementations, such as a better classifier. Also, the results from the classifier have no particular format, so we might need to use standardized data before elevating that data to a COVID-19 platform.

If COVIDUS is implemented, we need to allocate a lot of time and resources to designing and developing an AI-inference engine for it, a population-level AI-inference engine.

The Mobile App formats the survey answers before sending them to the Microservice. That bit may differ depending on the data type, which implies using an extra layer in the platform to format and possibly encode the data collected before sending it to a microservice.

We need to address the security and privacy concerns by authenticating the users. In [91], the topic of authenticating an individual with biometrics, specifically with ECG signals, is broadly studied, and biometric authentication can be one way to authenticate the PT; this approach needs to be investigated for COVIDMe.

We successfully implemented a DT for health in the context of COVID-19 using the reference model presented in [4]. The platform (COVIDMe) is designed so that we can replace or improve different components with relative ease and minimum effort. We made an argument regarding the accuracy of the AI-inference engine by explaining that accuracy depends mainly on the classifier's performance and created a design that supports different classifiers. We also outlined an architecture for a DT for health in the context of COVID-19, on which we can build and improve.

Chapter 5 presents the design and implementation of Cardio Twin, a PDT for well-being in the context of ischemic hearth diseases.

Chapter 5

5. Cardio Twin

This chapter discusses the design and implementation of a PDT for the human heart in the context of ischemic heart diseases (IHD), Cardio Twin, a PDT for detecting IHD. Cardio Twin aims to create a digital copy of a person's heart state using ECG and blood saturation data collected from the PT. The ECG signal is then used to detect non-myocardial and myocardial conditions using a Convolutional Neural Network (CNN) designed explicitly for this task and integrated into Cardio Twin. The "PTB Diagnostic ECG Database" [92], [93] from Physio Bank was used to train a CNN in detecting myocardial vs. non-myocardial states of the ECG signal. The database consists of 549 records from 290 subjects, of which 52 were healthy (i.e. non-myocardial). Each patient's data sample was divided into 2.5-second windows for training purposes. The CNN achieved an accuracy of 85.77% and took 4.8 seconds for each sample classification. These results indicate that technology can fully support demanding processes, such as Personal Digital Twins, even if running on the edge (i.e. a smartphone in this case).

Ischemic Heart disease (IHD) and stroke have been among the leading causes of death worldwide for several decades, particularly in populations with high-income levels. According to the World Health Organization, 2016 56.9 million deaths were recorded, of which 15.2 million were attributed to these two diseases [94], accounting for 26.7% of the total.

A Myocardial Infarction (MI) is a type of IHD in which the oxygen-rich blood flow to the heart muscle is interrupted, causing damage to the heart itself. Quite often, the victims of IHD do not realize they are about to suffer until it is too late, preventing them from getting proper help. Mortality, in this case, is related to a delay in treatment [87], which is why every minute counts from the moment the first symptoms appear until proper treatment is administered. In other words, early detection and reaction is critical to survival. Timely detection is not trivial as it requires the subject to be constantly monitored.

Consider the following scenario: a retired 62-year-old male who is a regular smoker is sitting in the dining room of his house, sharing the day with his family, when he suddenly suffers an MI. A call to the emergency services and hoping for the best is usually the only thing whoever is present can do for him at that point, and if he's alone, he will never receive any help. Continuous monitoring by a health professional is an excellent strategy to prevent this last scenario, but it is impossible to monitor all people at risk. However, the problem of a universal monitoring model remains unsolved. Hence, Cardio Twin is presented as a potential solution to this problem.

A PDT may use sensors and other data sources to create an exact virtual copy of a human being [5], [20]Cardio Twin, a PDT of the human heart, was designed using the design ecosystem for PDTs for well-being presented in Chapter 3. It is designed to automatically monitor and assist a person in an emergency event of IHD.

Time is critical when dealing with emergencies of IHD or Stroke; this is where the concept of Edge Computing comes into play. Satyanarayanan [95] calls edge computing devices "cloudlets" and lists the advantages: highly responsive cloud services, scalability via edge analytics, privacy-policy enforcements, and masking of cloud outages. Highly responsive cloud services refer to the almost non-existent latency that comes from the advantage of having most of the computation done in each cloudlet. Scalability via edge analytics frees the bandwidth load of the cloud by only transmitting the necessary data. Privacy is essential for Cardio Twin, and having it running on the edge contributes to it as the data is stored in the PT's smartphone. Thanks to edge computing, the analytics computing power required to monitor a subject continuously is always available at the subject's reach and service at any time, which translates to continuous monitoring. Edge computing enables the PT to control the information collected and processed (privacy). And "masking the cloud outages" pushes the service's availability into the subject's hands.

Much of the proposed DTs (PDTs included) for healthcare and well-being conform to a centralized architecture like the one described in [90], with “n” nodes collecting data and transmitting the data to a central structure (a server or the cloud) to process the data and send it back the results to the respective nodes like in figure 5-1. This architecture depends on a centralized “Data Analysis Structure.” The disadvantage of this system is that if this layer fails, it fails for all the patients. More examples of PDTs for healthcare can be found in [96], [97], [98]. A common characteristic of these systems is that they are oriented only to detect and store the subject's state, not to help them in an MI or even to prevent an IHD.

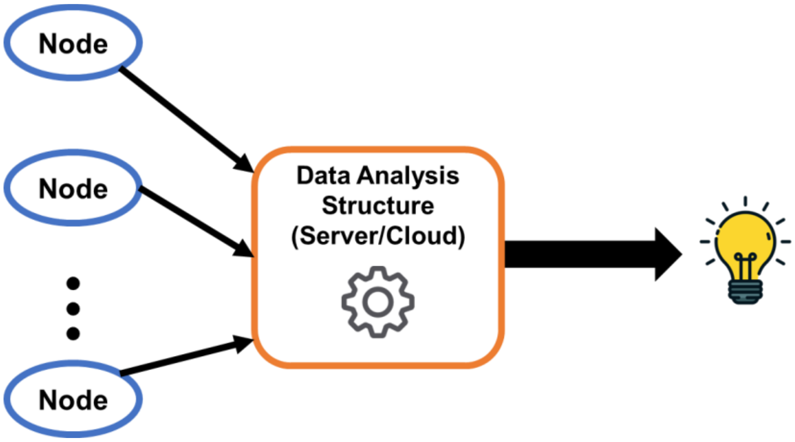


Figure 5-1. Typical configuration for data analysis of DTs for health existing in literature.

Smartphones have the intrinsic capacity to sense the user and their context by collecting data from the sensors integrated into the device. Then, it can also be enhanced by adding external sensors, taking advantage of its communication capabilities [99]. A plethora of ECG collecting devices with Bluetooth capabilities are available in the market, which means that these sensors can be easily attached to a smartphone; this makes them an ideal platform to deploy a PDT for healthcare, given their extensive use among the population of all ages. A smartphone ticks all the boxes to host Cardio Twin: edge device; recent models have increased computing capabilities, which makes them highly responsive, and the fact that they are mobile makes them easily accessible as they are within hand reach at almost all times. Deep learning classification models can run on modern smartphones.

5.1. Cardio Twin design

Cardio Twin collects data from wearable sensors; this data is processed to detect and help if the physical twin is suffering from an IHD or a Stroke. This platform uses internal sensors in edge devices such as smartphones and their capacity to pair with external sensors to collect biosignals through Bluetooth. Figure 5-2 shows the components of Cardio Twin. Machine learning interprets all the collected data and takes appropriate action by executing instruction pipelines. Cardio Twin communicates with external entities, such as hospitals, laboratories, or emergency services, through the smart services interface.

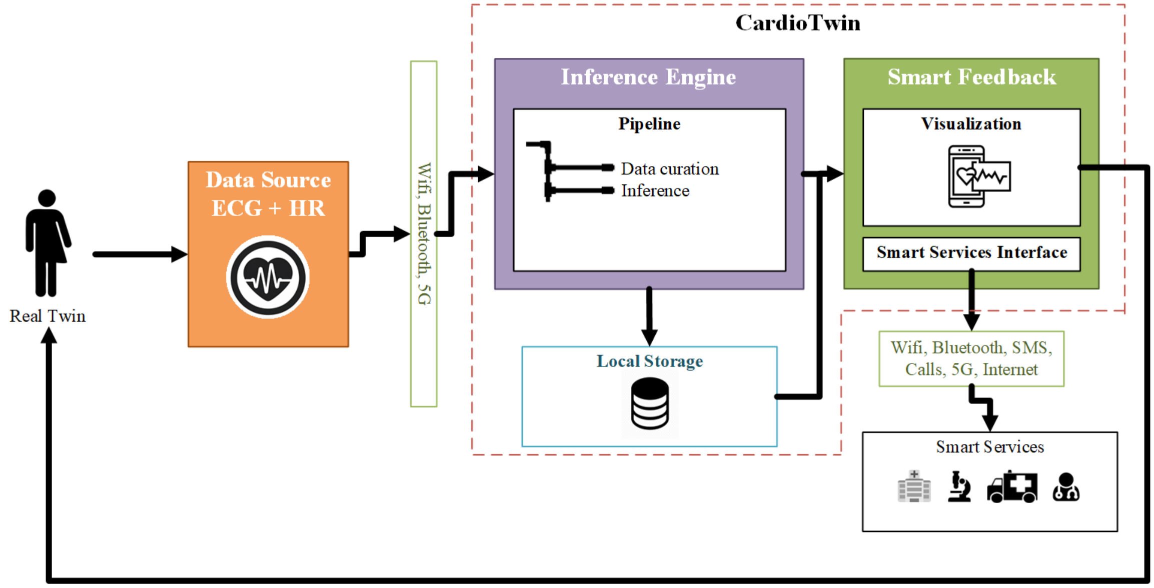


Figure 5-2. Cardio Twin architecture.

A smart feedback layer is considered in the design to render a representation of the human heart using augmented reality for the smart services, and the smartphone hosting Cardio Twin would provide on-screen feedback to the PT. Cardio Twin architecture is organized into three structures: Data Source, Inference Engine, and the Smart Feedback layer. Each one of the three core components is explained in detail next.

5.1.1. Data Source

Cardio Twin is designed to use ECG signals and blood saturation, the first collected with Kardia Mobile and the integrated heart rate sensor integrated into the smartphone (i.e. blood saturation). Details on the implementation will be provided next.

5.1.2. Inference Engine

All collected data, such as ECG and heart rate, is compiled into a single package and transmitted to the inference engine. This component is responsible for storing and recovering data from the local data storage module. The data is now available for analysis in the Inference Engine.

As mentioned, Cardio Twin's design is based on the design ecosystem presented in Chapter 3; however, the " pipeline " concept was added to the Inference Engine for Cardio Twin. This idea was inspired by the pipeline software architecture, in which a pipeline is formed by a series of sequentially executed pipes or processes that take the output of the previous pipe as an input. For Cardio Twin, a pipe is a software object that receives, processes and transmits data to the next object in the architecture. At the end of every pipeline, the data object is stored in the local storage. However, at the end of the execution of the last pipe, the inference pipe will transmit the output to the Smart Feedback component and store the inference in the local storage. The pipeline is executed as follows:

1. Data curation: This pipe collects and aggregates data from the ECG sensor in the data source and stores it in the smartphone storage system (Local Storage).
2. Inference Engine: This pipe recovers standardized data to analyze and discover “hidden” information. Cardio Twin uses Tensor Flow Lite models to classify this data and learn new information about the physical twin, which is useful for representing the physical twin’s heart. This information is then stored in Local Storage.

5.1.3. Smart Feedback

Each Cardio Twin must be able to provide feedback to its corresponding PT. To achieve this, the smart feedback module will either use the smartphone screen to give feedback to the user or use

Bluetooth connectivity, Wi-Fi/5G networks, and conventional messaging services to interface with the smart services outside Cardio Twin.

For illustrative purposes, imagine a physician who wants to examine a patient’s heart condition during a visit. The smart services interface will take care of the data transmission. With this feature, a patient could be virtually present in the doctor’s office regardless of distance, and the actual visit might not even be necessary.

The concept of smart service for healthcare and well-being is introduced in this architecture to describe a Cardio Twin compatible service that otherwise would require the intervention of the PT. For instance, if a cardio twin detects an anomaly in the heart, it could easily send the data from the past 48 hours to a laboratory using its communication capabilities. When necessary, it could also book an appointment with a physician. Healthcare providers offer smart services in Cardio Twin. In the scenario where a myocardial infarction is detected, a request for an ambulance could be sent to a Smart Service. This is accomplished through the Smart Services Interface. All these decisions are based on the inference obtained from the Inference pipe of the Inference Engine.

The next chapter presents a proof-of-concept implementation of Cardio Twin's inference engine to explore the challenges that may arise from implementing this architecture and, more specifically, the Inference Engine entirely on the edge.

5.2. Proof of concept implementation

At this point of maturity, it’s impossible to evaluate Cardio Twin in a real situation. To compensate for this, the proof-of-concept implementation is focused on the Inference Engine, more specifically, in the inference pipe. Pairing with external devices (such as sensors), accessing locally stored data, and communicating with other devices and services are intrinsic features of modern commercially available smartphones; these features are building blocks for the Smart Feedback and Local Storage components.

Thus, this implementation aimed to prove the Inference Engine's performance and Cardio Twin's viability in real scenarios since data analytics is usually done in a centralized server. In this case, the proposed implementation focuses only on monitoring the human heart to detect an abnormality in the heart ECG of the physical twin.

For instance, an event caused by an IHD would cause an abnormal ECG signal. The deep learning model for the inference pipe is a CNN trained to classify the first lead of a standard ECG (lead I). The training and test datasets were compiled based on “The PTB Diagnostic ECG Database” (PTBD), which can be found in Physio Bank [92] and described in [93] the process of compiling the train and test dataset was as follows: All the patients suffering from MI and healthy controls were identified from the PTBD database, 148 patients with MI and 52 healthy controls, for 200 patients.

The algorithm Pre-process ECG signals for training (Algorithm 5-1) is designed to prepare electrocardiogram (ECG) data for subsequent analysis, particularly for training machine learning models. The algorithm is executed as follows:

1. **Initialization**: The algorithm begins by defining key parameters, such as the sampling frequency (samplingFrequency) and the settings for the sliding windows (slidingWindowSettings). These parameters are crucial for standardizing the signal processing across all ECG data.
2. **Data Preparation**: A CSV file is initialized to store the processed data. The algorithm loads a directory containing ECG sessions (sessionsFolder), which includes data for multiple patients.
3. **Data Extraction and Processing**:
    a. For each patient folder within the sessions directory, the algorithm iterates over each ECG session file.
    b. It extracts essential information from each session file, including the patient's ID (patientId), the raw ECG signal (signal), and the patient’s condition (patientCondition).

4. **Signal Resampling**: The raw ECG signal is resampled to the defined samplingFrequency to ensure uniformity in the data analysis, accommodating the variability in recording devices or settings.
5. **Signal Detrending**: To enhance the quality of the signal for analysis, linear trends are removed from the entire signal. This step is critical for reducing potential biases in the data caused by baseline wander or other artifacts.
6. **Segmentation**: The detrended signal is then segmented into smaller portions using the sliding window technique, according to the slidingWindowSettings. This step is essential for creating manageable and analyzable units of data, each representing a specific time window of the ECG recording.
7. **Normalization and Data Formatting**:
    a. Each segment is scaled to a range between 0 and 1 to normalize the data, making it suitable for machine learning algorithms which perform better with scaled inputs.
    b. The scaled segment is then formatted into a comma-separated string.
8. **Data Point Creation**: A data point is created for each segment, incorporating the segment itself, the patient's condition, and the appropriate label, ready for analysis or model training.
9. **Data Storage**: Each data point is written to the CSV file, systematically compiling the preprocessed data.
10. **Completion**: Upon processing all sessions, the CSV file is closed, and a completion message is printed, indicating that all data sessions have been processed.

This comprehensive preprocessing approach ensures that the ECG data is uniformly formatted and enriched for effective use in training predictive models, enhancing the accuracy and reliability of any subsequent analyses or predictions based on this data.

Algorithm 5-1. Algorithm to pre-process ECG signals from the PTB Diagnostic ECG Database

```
Define samplingFrequency, slidingWindowSettings for ECG data processing
Initialize CSV file for output
sessionsFolder ← Load folder with ECG sessions
For each patientFolder in sessionsFolder
  For each ECG sessionFile in the patientFolder
    patientId, signal, patientCondition ← Extract data from sessionFile
    Resample signal to desired frequency
    Detrend the entire signal to remove linear trends
    segments ← Segment the detrended signal into sliding windows
    For each segment in segments
      Scale segment values between 0 and 1
      Convert segment to comma-separated string format
      dataPoint ← Use segment, patientCondition and label to create data point
      Write dataPoint to CSV file
Close CSV file
```

Using Algorithm 5-1, the first lead (lead I) of each ECG record of the 200 patients was recovered, scaled, and segmented to create 2.5-second samples fed to the model to ensure that at least one heartbeat was present in each sample. According to Edward Laskowski [100], a person in a resting state's heart rate ranges between 60 and 100 beats per minute. This implies that at the lowest heart rate, a healthy patient in a resting state will present a heartbeat per second. The time length for the segments for each sample was determined to guarantee that each segment had at least two heartbeats present. To properly evaluate the CNN model, the samples of 34 subjects out of the 200 subjects (healthy and non-healthy) were selected as test data. The rest of the samples of the remaining 166 subjects were used to build the training dataset for the CNN model. When working with neural networks, it is recommended that the values in the dataset are scaled to a range of 0-1. Because of that, each 2.5-second segment was scaled individually to fit in that range.

Algorithm 5-2. Algorithm for training and saving IHD detection model fitted with the dataset extracted from the PTB Diagnostic ECG Database.

```
Import dataset from dataSetPath

trainDataset, testDataset ← Perform a stratified split 70-30 to create train-test
datasets using original dataset

model ← Build CNN Model

trainedModel ← train model on trainDataset

evalResults ← evaluate trainedModel on testDataset

Plot evalResults

Save trainedModel
```

Once the training and test datasets were selected, the next step was to create a CNN model and train this model with the training dataset. Algorithm 5-2 was used to train the classifier; the algorithm is executed as follows:

1. **Initialization**: The algorithm begins by loading the pre-processed ECG dataset from the specified path (dataSetPath). This dataset contains ECG signals that have already been segmented, scaled, and labeled, ready for machine learning applications.
2. **Stratified Split**: Perform a stratified split on the dataset to divide it into training and testing subsets, typically using a 70-30 ratio. Stratification ensures that both subsets maintain a consistent distribution of classes, reflecting the prevalence of outcomes in the complete dataset. This is crucial for maintaining the model's ability to generalize across different data distributions.
3. **Build CNN Model**: Construct the CNN model. This includes setting up several convolutional layers to capture spatial and temporal features from the ECG signals. The layers may include ReLU activation functions to introduce non-linearity, pooling layers to reduce dimensionality, and fully connected layers at the end to make predictions.
4. **Train Model**: Train the CNN model on the training dataset. This process adjusts the model's weights to minimize the loss function, which measures the prediction errors. The model learns to correlate specific features of the ECG signals with the labeled outcomes, improving its accuracy over several iterations or epochs.

5. **Evaluate Model**: Once training is complete, evaluate the model's performance using the test dataset. This step assesses how well the model can predict new, unseen data, which is essential for its practical application. Evaluation metrics might include accuracy, precision, recall, and F1-score.
6. **Plot Results**: Finally, plot the evaluation results to visualize the model's performance across different metrics. This visualization typically includes plots of training and validation loss and accuracy over each epoch, providing insights into the model's learning progress and helping identify any issues like overfitting or underfitting.
7. **Completion**: Upon completing the evaluation, document the model's performance metrics for analysis. Save the trained model for future use, ensuring all results are securely stored and accessible for further testing or operational deployment. Confirm the successful execution of all processes and clean up any temporary data or resources utilized during the training session.

A heuristic approach was followed towards the construction of the CNN. Two main variables were selected: the number of convolution and pooling layers and the number of dense layers.

The highest accuracy was achieved with a model of 3 pairs of convolution-pooling layers, a flattened layer, two dense layers using relu as the activation function and one classification layer with a single neuron using sigmoid as the activation function as shown in Table 5-1. The model has an accuracy of 85.81%, a sensitivity of 86.29%, a precision of 95.6% and a specificity of 83.87%.

Table 5-1. Table detailing the architecture of the CNN model used to predict IHDs; this model was trained with the PTB Diagnostic ECG Database for training.

Layer	Parameters
Input Layer	input_shape = (625 x 1)
Conv1D	filters = 32 kernel_size = 4 activation = relu
MaxPooling1D	pool_size = 3
Conv1D	filters = 32 kernel_size = 4 activation = relu
MaxPooling1D	pool_size = 3
Conv1D	filters = 32 kernel_size = 4 activation = relu
MaxPooling1D	pool_size = 3
Flatten	
Dense	units = 128 activation = relu
Dense	units = 128 activation = relu
Dense	units = 1 activation = sigmoid

Figure 5-3 shows the architecture of the CNN model. The selected activation function was ReLU. Keras and TensorFlow are Python libraries that assist in designing and training deep-learning neural networks. Keras works on top of Tensor Flow.

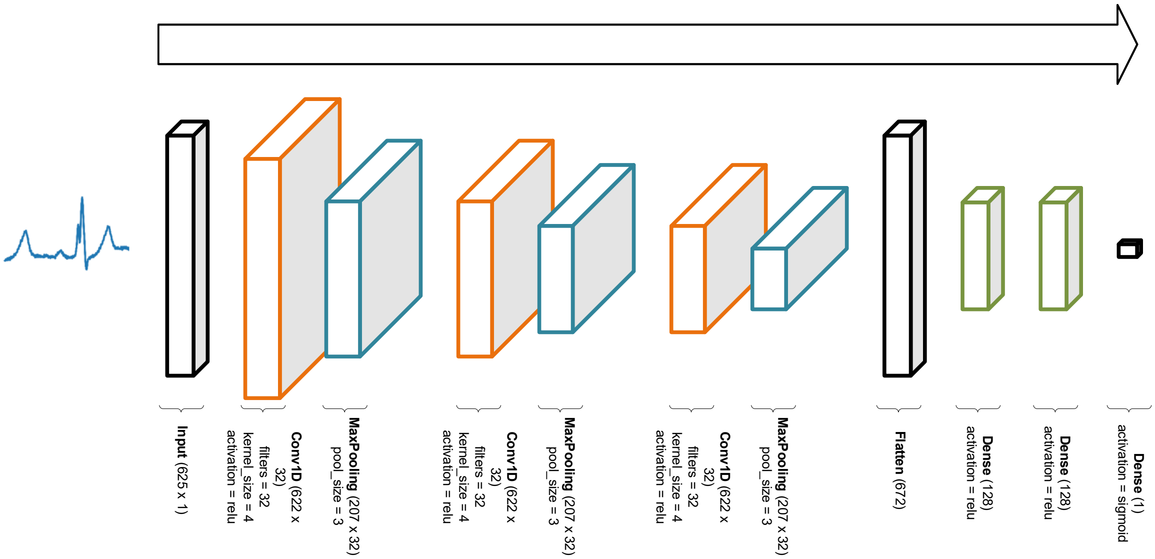
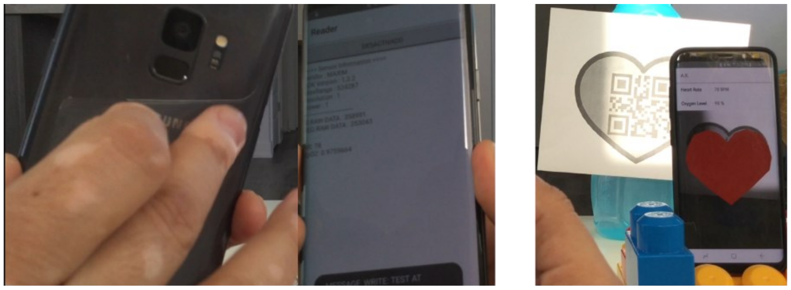


Figure 5-3. Architecture diagram of the CNN model used to predict IHDs; this CNN model was trained with the PTB Diagnostic ECG Database for training.

Cardio Twin is designed to run on the edge, hence, the most challenging element to implement in Cardio Twin is the Inference Engine since it requires the device to run classifiers in the device itself. The focus is on having an Inference Engine running on the edge that is versatile enough to run different models, allowing for optimization and improvements of the Machine Learning models running in the Inference Engine. This would allow for the models that run in the Inference Engine to be used to solve different tasks related to drawing inferences from the data collected from the human heart, thus detecting more than just an abnormal ECG signal. Keras is a Python library used to train deep neural network models, which is especially helpful when used to train convolutional neural networks. Keras has been used in recent years to tackle various classification problems. By running a Keras model, Cardio Twin can tackle as many classification problems as neural networks have tackled in the past, including deep learning models that have gained much attention from the academic community in recent years. Thus, running Keras models in this proof-of-concept is evidence of the viability of Cardio Twin architecture running entirely on edge.

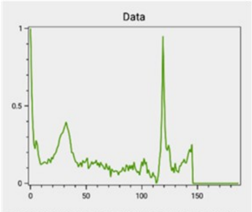
The proof-of-concept version of Cardio Twin was implemented in two stages. The first stage involved designing and training a CNN in Keras and obtaining a classifier for ECG segments into MI or non-MI segments. The second stage involved having the data analytics pipeline run the trained model and test its accuracy with the test dataset prepared in the first stage.

Figures 17c and 17d depict a prototypical visualization for Cardio Twin in which some samples are shown along with the predicted class (MI or non-MI). The classifier's output is shown on the “Predicted” label and compared to the “Expected” value. Cardio Twin visualization also allows you to see a graph of the ECG segment by pressing the “View Data” button. However, in this case, Cardio Twin was only tested against a fraction of the test set. Figure 5-4a shows the Cardio Twin running on a Galaxy S9 smartphone in a data collection state. Finally, Figure 5-4b shows a smartphone running proof of concept mobile application for a Cardio Twin “smart service"; the application renders a 3D model that follows the PTs real heart in real-time. The 3D model also reflects the data collected from the PT and the inferences that are drawn from the DCG signal (i.e., ischemic or not ischemic). The data collected in 17a is transmitted to the Smart Service in real-time, all while collecting the data from the PT and running the data through the Inference Engine.



a. Cardio Twin

b. Smart service proof of concept.



1, 0.6069411, 0.3841808, 0.2542373, 0.2235674, 0.2768362, 0.2534302, 0.1848265, 0.1533495, 0.1218725, 0.1251009, 0.1291364, 0.1372074, 0.1372074, 0.1323648, 0.1267151,

c. Cardio Twin. ECG data collection and rendering.

Expected Value: [1]

Predicted Value: [0.9996967000]

d. Cardio Twin. ECG-based IHD problem detection.

Figure 5-4. Cardio Twin prototype and, proof of concept smart service endpoint.

A second test was done using all 13420 ECG segments from the test set. To do this, the classification results and the time it took for the pipeline to process each sample are shown in Table 5-2.

Table 5-2. Performance of the TFLite model deployed with Cardio Twin to predict IHDs.

	Predicted Normal	**Predicted Abnormal**
Healthy	2213	1474
Myocardial	435	9298

Average time (ms)	4.8454545
Total Count	13420
Correct	11511
Incorrect	1909

Accuracy	85.77%
Precision	95.53%
Recall	86.32%

The accuracy of the trained model is 85.77%, which is quite close to the results obtained during the training phase in Keras. This suggests that other models implemented with the Keras framework can go through the same process without suffering significant alterations that could affect performance. Another interesting result is the relatively short time (4.84 ms) needed to classify a single segment. This classification time is short enough to allow real-time processing and classification of ECG segments.

Cardio Twin is a platform conceived as a twin of a human heart with the idea of detecting, preventing and reducing the risk of suffering heart diseases. The human heart is an extremely complex organ and cannot be fully replicated only by collecting ECG data, not even if it is real-time data. Therefore, it is necessary to identify and enable the platform to collect different data types from various sensors. The next step is to assemble all structures and modules, have the data fusion pipe collect data directly from the sensors, and increment the number of sensors so that

Cardio Twin can take full advantage of the sensing capacities of a modern smartphone. For instance, in contrast to ECG, phonocardiograms use clinical data to increase the model's accuracy in detecting heart problems, etc.

At the same time, work must be done on the multimodal interaction structure to enable better and richer visualizations of the Cardio Twin, which must be visualized on external displays and MR devices.

At this point, the Cardio Twin focuses on detecting a problem and acts accordingly to help the physical twin. This is one way to approach this problem, but in the future, the platform will help prevent IHD and Stroke by reducing the risk factors associated with these diseases. The suggested approach is to apply persuasive computing to promote healthy habits in the physical twin, thus reducing risk factors such as alcohol consumption, smoking, sedentary lifestyle, etc. The AI inference engine will be key in achieving this.

The platform's proof-of-concept implementation was successful even though only ECG data was used for the data Inference pipe; a similar process can be followed to integrate other types of data, which opens the possibility of using Cardio Twin. On the other hand, whenever a model is not accurate enough, the design of Cardio Twin allows the replacement or improvement of any model in the data analytics pipe. The architecture is notable for implementing the Inference Engine on edge; this pushes the “brains” of the platform entirely to the edge, which is a relevant result. In summary, a Digital Twin of the human heart running entirely on the edge is a viable idea that creates new research opportunities to exploit.

Chapter 6

6. Conclusions and Future Work

The present research work explores the integration of personal digital twins for health. There are several use cases and applications for it in the area of mental and physical well-being and physical health assessment and monitoring; these applications are possible thanks to the use of Machine Learning models that provide the PDTs with the "inference" capabilities useful for the evaluation and decision-making. Deploying a PDT for well-being is a novel idea on its own, but at the same time, it raises interesting questions. The first is the "How?" at the core of the present research: a design ecosystem to design personal digital twins for well-being. However scarce, there are efforts to implement PDTs for health, most of which are directed at addressing a particular use case rather than designing with standardization, scalability or generalization in mind. The ecosystem presented in this document is a novel concept in healthcare, leveraging the power of virtual reality and advanced data analytics (i.e. Machine Learning). The concept extends beyond traditional telemedicine, offering a more immersive, interactive, and personalized healthcare experience, making the ideal of precision medicine attainable. This approach could significantly improve patient outcomes and overall well-being management.

The successful implementation of PDTs requires the seamless integration of diverse and complex data streams. This includes electronic health records, biometric data from wearables, and environmental data. Ensuring the privacy and security of this sensitive data is paramount

AI plays a crucial role in the functionality of PDTs, analyzing vast amounts of data to provide insights and predictive models for individual health. Integrating AI with PDTs could lead to more accurate diagnoses, personalized treatment plans, and even predictive healthcare analytics, opening the door to unexplored research lines in eHealth.

This paradigm shift also impacts healthcare professionals. Accessing a patient's digital twin could aid diagnosis, treatment planning, and remote patient monitoring. However, it also necessitates new skills and training for healthcare providers to utilize the latest technology effectively.

Deploying PDTs for well-being raises ethical questions around data ownership, consent, and the digital divide. Ensuring equitable access to this technology is crucial to avoid exacerbating healthcare access disparities.

Integrating a design ecosystem for the design of Personal Digital twins for well-being marks a significant milestone in the evolution of healthcare. This novel approach may open the doors to new research lines and set a baseline from which new research on the topic may start. It also represents a paradigm shift and may open the door for more accessible and preventive healthcare. It also opens new avenues for patient engagement and personalized care.

While the opportunities are vast, the challenges are equally significant. Ensuring data integrity, privacy, and security is crucial. Additionally, successfully adopting this technology requires addressing infrastructural, educational, and ethical barriers.

PDTs hold immense promise in preventive medicine. By offering real-time health data collection and predictive insights, they can enable early intervention and more effective management of chronic diseases.

Realizing the full potential of PDTs for well-being requires a collaborative effort between researchers, engineers, healthcare providers, policymakers, and patients; it necessitates a multidisciplinary approach. As this technology progresses, it's imperative to ensure that it is accessible and beneficial to all. This calls for strategies to bridge the digital divide and make digital healthcare access a tool for reducing, not exacerbating, health access disparities.

6.1. Future Work

Future work should focus on creating standards and protocols for health data transmission. This interoperability is crucial for integrating disparate healthcare systems and devices, facilitating a more cohesive and efficient digital healthcare delivery environment.

As health data becomes increasingly digitized and compiled, robust security measures are essential. Research should focus on enabling secure and strong communication of health data among stakeholders; research may leverage advanced encryption methods, blockchain technology for data integrity, and secure authentication mechanisms to protect sensitive health information.

Future research should explore the integration of additional biometric sensors and health indicators into virtual environment-oriented PDTs. For this thesis, the list of health indicators may include genetic data, environmental factors, and social determinants of health, providing a more holistic view of individual well-being.

There is a need for sophisticated AI algorithms capable of handling the complexity and volume of data within PDTs. These algorithms should analyze current health data and predict future health trajectories, aiding in preventive medicine and personalized health interventions.

Future work must also include developing ethical frameworks and policies to govern the use of PDTs for well-being. This includes addressing consent, data ownership, privacy, and equitable access to these technologies.

Finally, the TMW design ecosystem requires a comprehensive evaluation framework to assess its effectiveness. Such a framework is crucial for understanding and measuring the impact of TMW-based PDTs on various well-being interventions, including mental health, rehabilitation, and nutrition. The TMW ecosystem incorporates QoE-based feedback, allowing for the integration of techniques based on the Physical Twin's perception. In Human-Centered Computing, these approaches are widely used and include surveys like Perceived Ease of Use, focus groups, and interviews; in addition to direct measurements of outcomes through quantitative analysis.

7. References

[1] "Health and Well-Being." Accessed: Feb. 09, 2024. [Online]. Available: https://www.who.int/data/gho/data/major-themes/health-and-well-being

[2] N. Bagaria, F. Laamarti, H. F. Badawi, A. Albraikan, R. A. Martinez Velazquez, and A. El Saddik, "Health 4.0: Digital Twins for Health and Well-Being," in Connected Health in Smart Cities, Cham: Springer International Publishing, 2020, pp. 143–152. doi: 10.1007/978-3-030-27844-1_7.

[3] M. N. Kamel Boulos and P. Zhang, "Digital Twins: From Personalised Medicine to Precision Public Health," Journal of Personalized Medicine 2021, Vol. 11, Page 745, vol. 11, no. 8, p. 745, Jul. 2021, doi: 10.3390/JPM11080745.

[4] A. El Saddik et al., "Dtwins: A Digital Twins Ecosystem For Health And Well-Being," IEEE COMSOC MMTC Communications -Frontiers, vol. 14, pp. 39–46, 2019.

[5] A. El Saddik, "Digital Twins: The Convergence of Multimedia Technologies," IEEE MultiMedia, vol. 25, no. 2, pp. 87–92, Apr. 2018, doi: 10.1109/MMUL.2018.023121167.

[6] W. Y. B. Lim et al., "Realizing the Metaverse with Edge Intelligence: A Match Made in Heaven," IEEE Wirel Commun, pp. 1–9, 2022, doi: 10.1109/MWC.018.2100716.

[7] "Constitution of the World Health Organization." Accessed: Aug. 29, 2023. [Online]. Available: https://www.who.int/about/governance/constitution

[8] J. Raibley, "Well-Being and the Priority of Values," Soc Theory Pract, vol. 36, no. 4, pp. 593–620, 2010, doi: 10.5840/soctheorpract201036432.

[9] I. Vlase and R. Sieber, "Narrating well-being in the context of precarious prosperity: An account of agency framed by culturally embedded happiness and gender beliefs," The

European journal of women’s studies, vol. 23, no. 2, pp. 185–199, 2016, doi: 10.1177/1350506815580399.

[10] J. Kabat-Zinn, “A Study in Happiness—Meditation, the Brain, and the Immune System,” Mindfulness (N Y), vol. 9, no. 5, pp. 1664–1667, Oct. 2018, doi: 10.1007/s12671-018-0991-3.

[11] E. M. V. Reiche, S. O. V. Nunes, and H. K. Morimoto, “Stress, depression, the immune system, and cancer,” Lancet Oncol, vol. 5, no. 10, pp. 617–625, Oct. 2004, doi: 10.1016/S1470-2045(04)01597-9.

[12] R. Veenhoven, “Healthy happiness: effects of happiness on physical health and the consequences for preventive health care,” J Happiness Stud, vol. 9, no. 3, pp. 449–469, Sep. 2008, doi: 10.1007/s10902-006-9042-1.

[13] J. K. Kiecolt-Glaser, L. McGuire, T. F. Robles, and R. Glaser, “Psychoneuroimmunology: Psychological influences on immune function and health.,” J Consult Clin Psychol, vol. 70, no. 3, pp. 537–547, 2002, doi: 10.1037/0022-006X.70.3.537.

[14] D. Hillel. Gelernter, Mirror worlds, or, The day software puts the universe in a shoebox-- : how it will happen and what it will mean. New York: Oxford University Press, 1991.

[15] M. W. Grieves, “Virtually Intelligent Product Systems: Digital and Physical Twins,” in Complex Systems Engineering: Theory and Practice, Reston, VA: American Institute of Aeronautics and Astronautics, Inc., 2019, pp. 175–200. doi: 10.2514/5.9781624105654.0175.0200.

[16] M. Grieves and J. Vickers, “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems,” Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches, pp. 85–113, Jan. 2017, doi: 10.1007/978-3-319-38756-7_4.

[17] R. Martinez-Velazquez, R. Gamez, and A. El Saddik, “Cardio Twin: A Digital Twin of the human heart running on the edge,” in 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE, Jun. 2019, pp. 1–6. doi: 10.1109/MeMeA.2019.8802162.

[18] S. Ishaque, A. Rueda, B. Nguyen, N. Khan, and S. Krishnan, “Physiological Signal Analysis and Classification of Stress from Virtual Reality Video Game,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2020-July, pp. 867–870, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176110.

[19] Z. Ahmad and N. M. Khan, “Multi-level Stress Assessment Using Multi-domain Fusion of ECG Signal,” Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2020-July, pp. 4518–4521, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176590.

[20] K. M. Alam and A. El Saddik, “C2PS: A Digital Twin Architecture Reference Model for the Cloud-Based Cyber-Physical Systems,” IEEE Access, vol. 5, pp. 2050–2062, 2017, doi: 10.1109/ACCESS.2017.2657006.

[21] P. Barbiero, R. Viñas Torné, and P. Lió, “Graph Representation Forecasting of Patient’s Medical Conditions: Toward a Digital Twin,” Front Genet, vol. 12, p. 1289, Sep. 2021, doi: 10.3389/FGENE.2021.652907/BIBTEX.

[22] A. Croatti, M. Gabellini, S. Montagna, and A. Ricci, “On the Integration of Agents and Digital Twins in Healthcare,” J Med Syst, vol. 44, no. 9, pp. 1–8, Sep. 2020, doi: 10.1007/S10916-020-01623-5/FIGURES/2.

[23] A. Karakra, F. Fontanili, E. Lamine, and J. Lamothe, “HospiT’Win: A predictive simulation-based digital twin for patients pathways in hospital,” 2019 IEEE EMBS

International Conference on Biomedical and Health Informatics, BHI 2019 - Proceedings, May 2019, doi: 10.1109/BHI.2019.8834534.

[24] A. Javaux et al., “A mixed-reality surgical trainer with comprehensive sensing for fetal laser minimally invasive surgery,” Int J Comput Assist Radiol Surg, vol. 13, no. 12, pp. 1949–1957, Dec. 2018, doi: 10.1007/s11548-018-1822-7.

[25] S. F. Hardon, A. Kooijmans, R. Horeman, M. van der Elst, A. L. A. Bloemendaal, and T. Horeman, “Validation of the portable virtual reality training system for robotic surgery (PoLaRS): a randomized controlled trial,” Surg Endosc, vol. 36, no. 7, pp. 5282–5292, Jul. 2022, doi: 10.1007/s00464-021-08906-z.

[26] “Da Vinci Xi Robotic Surgical System | Intuitive.” Accessed: Apr. 02, 2024. [Online]. Available: https://www.intuitive.com/en-us/products-and-services/da-vinci/xi

[27] M. W. Zackoff et al., “Immersive Virtual Reality Onboarding using a Digital Twin for a New Clinical Space Expansion: A Novel Approach to Large-Scale Training for Health Care Providers,” J Pediatr, vol. 252, pp. 7-10.e3, Jan. 2023, doi: 10.1016/j.jpeds.2022.07.031.

[28] D. Takagi et al., “Effects of dental students’ training using immersive virtual reality technology for home dental practice,” Educ Gerontol, vol. 45, no. 11, pp. 670–680, Nov. 2019, doi: 10.1080/03601277.2019.1686284.

[29] V. Gagnon Shaigetz et al., “An Immersive and Interactive Platform for Cognitive Assessment and Rehabilitation (bWell): Design and Iterative Development Process,” JMIR Rehabil Assist Technol, vol. 8, no. 4, p. e26629, Nov. 2021, doi: 10.2196/26629.

[30] C. Hanshans, B. Maisch, J. Zauner, M. M. R. Faust, L. M. Bröll, and S. Karch, “Virtual Therapeutics – Requirements to deliver value with virtual reality and biofeedback

applications for alcohol addiction therapy," Current Directions in Biomedical Engineering, vol. 7, no. 2, pp. 81–84, Oct. 2021, doi: 10.1515/cdbme-2021-2021.

[31] U. Rajashekhar and N. Neelappa, "Development of Automated BCI System to Assist the Physically Challenged Person Through Audio Announcement With Help of EEG Signal," WSEAS TRANSACTIONS ON SYSTEMS AND CONTROL, vol. 16, pp. 302–314, May 2021, doi: 10.37394/23203.2021.16.26.

[32] J. Qu, L. Cui, W. Guo, X. Ren, and L. Bu, "The Effects of a Virtual Reality Rehabilitation Task on Elderly Subjects: An Experimental Study Using Multimodal Data," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 1684–1692, 2022, doi: 10.1109/TNSRE.2022.3183686.

[33] Y. Chen et al., "Digital-Twin-Based Patient Evaluation during Stroke Rehabilitation," ICCPS 2023 - Proceedings of the 2023 ACM/IEEE 14th International Conference on Cyber-Physical Systems with CPS-IoT Week 2023, pp. 22–33, May 2023, doi: 10.1145/3576841.3585923.

[34] L. F. Rivera, M. Jiménez, P. Angara, N. M. Villegas, G. Tamura, and H. A. Müller, "Towards Continuous Monitoring in Personalized Healthcare through Digital Twins," 2019, doi: 10.5555/3370272.

[35] A. Costanzo, A. Faro, D. Giordano, and C. Pino, "Mobile cyber physical systems for health care: Functions, ambient ontology and e-diagnostics," 2016 13th IEEE Annual Consumer Communications and Networking Conference, CCNC 2016, pp. 972–975, Mar. 2016, doi: 10.1109/CCNC.2016.7444920.

[36] S. Khan, S. Ullah, H. U. Khan, and I. U. Rehman, "Digital-Twins-Based Internet of Robotic Things for Remote Health Monitoring of COVID-19 Patients," IEEE Internet Things J, vol. 10, no. 18, pp. 16087–16098, Sep. 2023, doi: 10.1109/JIOT.2023.3267171.

[37] Y. Liu et al., “A Novel Cloud-Based Framework for the Elderly Healthcare Services Using Digital Twin,” IEEE Access, vol. 7, pp. 49088–49101, 2019, doi: 10.1109/ACCESS.2019.2909828.

[38] I. Isiksalan, “MedicaLog: Transforming Healthcare with AI-Powered Health Data Platform,” BioCAS 2023 - 2023 IEEE Biomedical Circuits and Systems Conference, Conference Proceedings, 2023, doi: 10.1109/BIOCAS58349.2023.10388828.

[39] L. Abirami and J. Karthikeyan, “Digital Twin-Based Healthcare System (DTHS) for Earlier Parkinson Disease Identification and Diagnosis Using Optimized Fuzzy Based k-Nearest Neighbor Classifier Model,” IEEE Access, vol. 11, pp. 96661–96672, 2023, doi: 10.1109/ACCESS.2023.3312278.

[40] Y. Chahal, R. Tokas, and K. Sharma, “Smart Solution Using Digital Twin and IoT for Diabetic Retinopathy,” 2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023, 2023, doi: 10.1109/ICCCNT56998.2023.10307620.

[41] K. Tang, K. Gerling, and L. Geurts, “Virtual Feed: Design and Evaluation of a Virtual Reality Simulation Addressing the Lived Experience of Breastfeeding,” in CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM, Apr. 2022, pp. 1–17. doi: 10.1145/3491102.3517620.

[42] F. Grieger, H. Klapperich, and M. Hassenzahl, “Trash It, Punch It, Burn It – Using Virtual Reality to Support Coping with Negative Thoughts,” in Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM, May 2021, pp. 1–6. doi: 10.1145/3411763.3451738.

[43] N. Wagener, J. Niess, Y. Rogers, and J. Schöning, “Mood Worlds: A Virtual Environment for Autonomous Emotional Expression,” Conference on Human Factors in Computing Systems - Proceedings, p. 16, Apr. 2022, doi: 10.1145/3491102.3501861.

[44] T. Huber, S. Mertes, S. Rangelova, S. Flutura, and E. Andre, “Dynamic Difficulty Adjustment in Virtual Reality Exergames through Experience-driven Procedural Content Generation,” in 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, Dec. 2021, pp. 1–8. doi: 10.1109/SSCI50451.2021.9660086.

[45] B. Subramanian, J. Kim, M. Maray, and A. Paul, “Digital Twin Model: A Real-Time Emotion Recognition System for Personalized Healthcare,” IEEE Access, vol. 10, pp. 81155–81165, 2022, doi: 10.1109/ACCESS.2022.3193941.

[46] O. Shaer, J. Otiono, Z. Qian, A. Seals, and O. Nov, “Remote evaluation of augmented reality interaction with personal health information,” Front Comput Sci, vol. 4, Aug. 2022, doi: 10.3389/fcomp.2022.934694.

[47] M. S. Khan, V. Charissis, and S. Sakellariou, “Exploring the Development Requirements for Virtual Reality Gait Analysis,” Multimodal Technologies and Interaction, vol. 3, no. 2, p. 24, Apr. 2019, doi: 10.3390/mti3020024.

[48] D. Mourtzis, J. Angelopoulos, N. Panopoulos, and D. Kardamakis, “A Smart IoT Platform for Oncology Patient Diagnosis based on AI: Towards the Human Digital Twin,” Procedia CIRP, vol. 104, pp. 1686–1691, Jan. 2021, doi: 10.1016/J.PROCIR.2021.11.284.

[49] T. Zhao, J. Liu, Y. Wang, H. Liu, and Y. Chen, “PPG-based Finger-level Gesture Recognition Leveraging Wearables,” in IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, IEEE, Apr. 2018, pp. 1457–1465. doi: 10.1109/INFOCOM.2018.8486006.

[50] M. McMahon and M. Schukat, “A low-Cost, Open-Source, BCI- VR Game Control Development Environment Prototype for Game Based Neurorehabilitation,” in 2018 IEEE Games, Entertainment, Media Conference (GEM), IEEE, Aug. 2018, pp. 1–9. doi: 10.1109/GEM.2018.8516468.

[51] C. Versek et al., “Portable System for Neuro-Optical Diagnostics Using Virtual Reality Display,” Mil Med, vol. 184, no. Supplement_1, pp. 584–592, Mar. 2019, doi: 10.1093/milmed/usy286.

[52] E. R. Sykes, G. Tofighi, Y. Wang, D. Ganguly, A. Janwaar, and A. Singh, “On the Design of Therapy Tracker: A Cross-platform Medication Management Application with Augmented Reality and Machine Learning,” 2022, pp. 29–48. doi: 10.1007/978-3-030-89906-6_3.

[53] K. Klinker, M. Wiesche, and H. Krcmar, “Digital Transformation in Health Care: Augmented Reality for Hands-Free Service Innovation,” Information Systems Frontiers, vol. 22, no. 6, pp. 1419–1431, Dec. 2020, doi: 10.1007/s10796-019-09937-7.

[54] J. J. Lee, M. Klepcha, M. Wong, P. N. Dang, S. S. Sadrameli, and G. W. Britz, “The First Pilot Study of an Interactive, 360° Augmented Reality Visualization Platform for Neurosurgical Patient Education: A Case Series,” Operative Neurosurgery, vol. 23, no. 1, pp. 53–59, Jul. 2022, doi: 10.1227/ons.0000000000000186.

[55] N. Wiener, “Cybernetics; Or, Control and Communication in the Animal and the Machine,” 2D Ed., Cambridge, Mass. : M.I.T. Press, 1961, 1961, ch. IV.

[56] C. E. Shannon, “A Mathematical Theory of Communication,” Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/J.1538-7305.1948.TB01338.X.

[57] T. Ege and K. Yanai, “Simultaneous estimation of food categories and calories with multi-task CNN,” in 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), IEEE, May 2017, pp. 198–201. doi: 10.23919/MVA.2017.7986835.

[58] K. Okamoto and K. Yanai, “An Automatic Calorie Estimation System of Food Images on a Smartphone,” in Proceedings of the 2nd International Workshop on Multimedia Assisted

Dietary Management, New York, NY, USA: ACM, Oct. 2016, pp. 63–70. doi: 10.1145/2986035.2986040.

[59] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, “Smartwatch-based activity recognition: A machine learning approach,” in 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, Feb. 2016, pp. 426–429. doi: 10.1109/BHI.2016.7455925.

[60] A. Sano et al., “Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones,” in 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), IEEE, Jun. 2015, pp. 1–6. doi: 10.1109/BSN.2015.7299420.

[61] M. Salath et al., “COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation,” Swiss Med Wkly, Mar. 2020, doi: 10.4414/smw.2020.20225.

[62] C. R. Wells et al., “Optimal COVID-19 quarantine and testing strategies,” Nat Commun, vol. 12, no. 1, p. 356, Dec. 2021, doi: 10.1038/s41467-020-20742-8.

[63] J. Summers et al., “Potential lessons from the Taiwan and New Zealand health responses to the COVID-19 pandemic,” Lancet Reg Health West Pac, vol. 4, p. 100044, Nov. 2020, doi: 10.1016/j.lanwpc.2020.100044.

[64] “Nucleic Acid Amplification Tests (NAATs) | CDC.” Accessed: Dec. 04, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/lab/naats.html

[65] “Interim Guidance for Antigen Testing for SARS-CoV-2 | CDC.” Accessed: Dec. 04, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html

[66] "Test for Current Infection | CDC." Accessed: Dec. 04, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/testing/diagnostic-testing.html

[67] J. Elflein, "COVID-19 testing rate by country | Statista." Accessed: Apr. 29, 2021. [Online]. Available: https://www.statista.com/statistics/1104645/covid19-testing-rate-select-countries-worldwide/

[68] D. Dowdy and G. D'Souza, "COVID-19 Testing: Understanding the 'Percent Positive' | Johns Hopkins Bloomberg School of Public Health." Accessed: Nov. 29, 2021. [Online]. Available: https://publichealth.jhu.edu/2020/covid-19-testing-understanding-the-percent-positive

[69] J. H. U. & Medicine, "How Does Testing in the U.S. Compare to Other Countries? - Johns Hopkins Coronavirus Resource Center," Testing Hub - Data Visualizations. Accessed: Nov. 29, 2021. [Online]. Available: https://coronavirus.jhu.edu/testing/international-comparison

[70] Y. LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," Neural Comput, vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

[71] D. P. Fan et al., "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images," IEEE Trans Med Imaging, vol. 39, no. 8, pp. 2626–2637, Aug. 2020, doi: 10.1109/TMI.2020.2996645.

[72] X. Wang et al., "A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization from Chest CT," IEEE Trans Med Imaging, vol. 39, no. 8, pp. 2615–2625, Aug. 2020, doi: 10.1109/TMI.2020.2995965.

[73] E. F. Ohata et al., "Automatic detection of COVID-19 infection using chest X-ray images through transfer learning," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 1, pp. 239–248, Jan. 2021, doi: 10.1109/JAS.2020.1003393.

[74] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," Comput Biol Med, vol. 121, p. 103792, Jun. 2020, doi: 10.1016/J.COMPBIOMED.2020.103792.

[75] M. J. Horry et al., "COVID-19 Detection through Transfer Learning Using Multimodal Imaging Data," IEEE Access, vol. 8, pp. 149808–149824, 2020, doi: 10.1109/ACCESS.2020.3016780.

[76] S. Hu et al., "Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification from CT Images," IEEE Access, vol. 8, pp. 118869–118883, 2020, doi: 10.1109/ACCESS.2020.3005510.

[77] T. K. Dash, S. Mishra, G. Panda, and S. C. Satapathy, "Detection of COVID-19 from speech signal using bio-inspired based cepstral features.," Pattern Recognit, vol. 117, p. 107999, Sep. 2021, doi: 10.1016/j.patcog.2021.107999.

[78] A. Tena, F. Clarià, and F. Solsona, "Automated detection of COVID-19 cough," Biomed Signal Process Control, vol. 71, p. 103175, Jan. 2022, doi: 10.1016/j.bspc.2021.103175.

[79] E. A. Mohammed, M. Keyhani, A. Sanati-Nezhad, S. H. Hejazi, and B. H. Far, "An ensemble learning approach to digital corona virus preliminary screening from cough sounds," Scientific Reports 2021 11:1, vol. 11, no. 1, pp. 1–11, Jul. 2021, doi: 10.1038/s41598-021-95042-2.

[80] T. Xia, Y. Q. Fu, N. Jin, P. Chazot, P. Angelov, and R. Jiang, "AI-enabled Microscopic Blood Analysis for Microfluidic COVID-19 Hematology," Proceedings - 2020 5th International Conference on Computational Intelligence and Applications, ICCIA 2020, pp. 98–102, Jun. 2020, doi: 10.1109/ICCIA49625.2020.00026.

[81] V. A. de Freitas Barbosa et al., “Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests,” Research on Biomedical Engineering 2021, pp. 1–18, Jan. 2021, doi: 10.1007/S42600-020-00112-5.

[82] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, “Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study,” Journal of Medical Systems 2020 44:8, vol. 44, no. 8, pp. 1–12, Jul. 2020, doi: 10.1007/S10916-020-01597-4.

[83] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, “Ensemble learning model for diagnosing COVID-19 from routine blood tests,” Inform Med Unlocked, vol. 21, p. 100449, Jan. 2020, doi: 10.1016/J.IMU.2020.100449.

[84] R. Martinez-Velazquez, D. P. Tobón V., A. Sanchez, A. El Saddik, and E. Petriu, “A Machine Learning Approach as an Aid for Early COVID-19 Detection,” Sensors 2021, Vol. 21, Page 4202, vol. 21, no. 12, p. 4202, Jun. 2021, doi: 10.3390/S21124202.

[85] R. Martinez-Velazquez, F. Ceballos, A. Sanchez, A. El Saddik, and E. Petriu, “COVIDMe: a digital twin for COVID-19 self-assessment and detection,” in Digital Twin for Healthcare, Elsevier, 2023, pp. 137–156. doi: 10.1016/B978-0-32-399163-6.00012-3.

[86] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of machine learning research, 2011.

[87] L. Breiman, Classification and Regression Trees, First edit. Boca Raton, FL: CRC Press, 2017.

[88] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, Dec. 2015. Accessed: May 05, 2021. [Online]. Available: https://arxiv.org/abs/1412.6980v9

[89] “Symptoms of COVID-19 | CDC.” Accessed: Nov. 27, 2023. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[90] J. Bergstra, J. B. Ca, and Y. B. Ca, “Random Search for Hyper-Parameter Optimization Yoshua Bengio,” 2012. Accessed: May 05, 2021. [Online]. Available: http://scikit-learn.sourceforge.net.

[91] J. S. Arteaga Falconi and A. El Saddik, “Security with ECG Biometrics BT - Advanced Methods for Human Biometrics,” N. Derbel and O. Kanoun, Eds., Cham: Springer International Publishing, 2021, pp. 65–79. doi: 10.1007/978-3-030-81982-8_4.

[92] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet,” Circulation, vol. 101, no. 23, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.

[93] R. Bousseljot, D. Kreiseler, and A. Schnabel, “Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet,” Biomedizinische Technik/Biomedical Engineering, pp. 317–318, Jul. 2009, doi: 10.1515/bmte.1995.40.s1.317.

[94] “The top 10 causes of death.” Accessed: Nov. 26, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[95] M. Satyanarayanan, “The emergence of edge computing,” Computer (Long Beach Calif), vol. 50, no. 1, pp. 30–39, Jan. 2017, doi: 10.1109/MC.2017.9.

[96] D. Gil, A. Hernandez-Sabate, D. Castells-Rufas, and J. Carrabina, “CYBERH: Cyber-physical systems in health for personalized assistance,” Proceedings - 2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017, pp. 373–376, Nov. 2018, doi: 10.1109/SYNASC.2017.00067.

[97] Z. Fu, C. Guo, S. Ren, Y. Ou, and L. Sha, “Modeling and Integrating Human Interaction Assumptions in Medical Cyber-Physical System Design,” Proc IEEE Symp Comput Based Med Syst, vol. 2017-June, pp. 373–378, Nov. 2017, doi: 10.1109/CBMS.2017.50.

[98] T. Shah et al., “Remote health care cyber-physical system: quality of service (QoS) challenges and opportunities,” IET Cyber-Physical Systems: Theory & Applications, vol. 1, no. 1, pp. 40–48, Dec. 2016, doi: 10.1049/IET-CPS.2016.0023.

[99] I. R. Félix, L. A. Castro, L. F. Rodríguez, and É. C. Ruiz, “Mobile phone sensing: Current trends and challenges,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9454, pp. 369–374, 2015, doi: 10.1007/978-3-319-26401-1_34/COVER.

[100] “Heart rate: What’s normal? - Mayo Clinic.” Accessed: Nov. 26, 2023. [Online]. Available: https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979