11-2023

A NOVEL MULTI-MODEL PATIENT SIMILARITY NETWORK DRIVEN BY FEDERATED DATA QUALITY AND RESOURCE PROFILING

Alramzana Nujum Navaz

UAEU

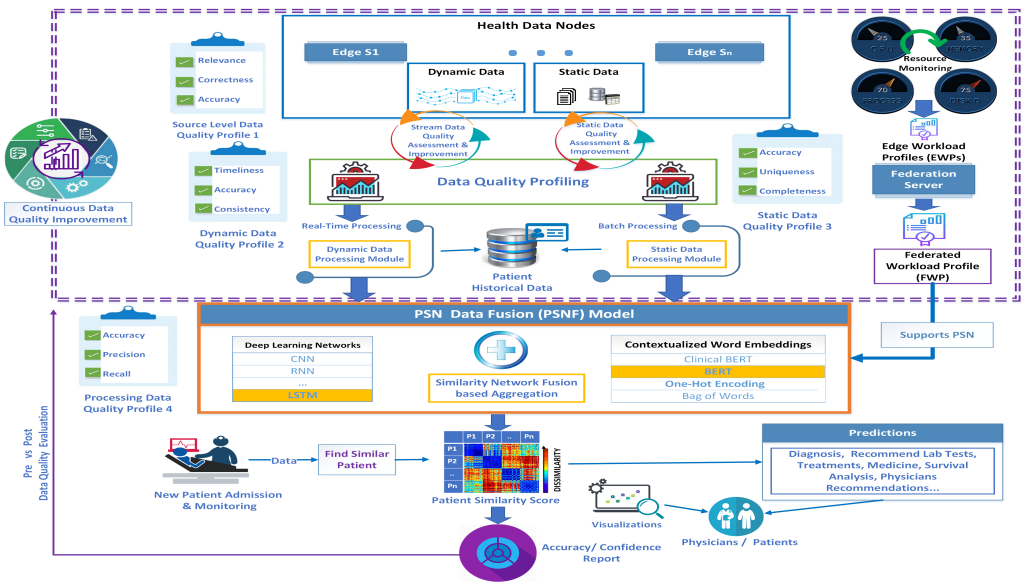
جامعة الإمارات العربية المتحدة
United Arab Emirates University

DOCTORATE DISSERTATION NO. 2023: 66

College of Information Technology

A NOVEL MULTI-MODEL PATIENT SIMILARITY NETWORK DRIVEN BY FEDERATED DATA QUALITY AND RESOURCE PROFILING

Alramzana Nujum Navaz



November 2023

United Arab Emirates University

College of Information Technology

A NOVEL MULTI-MODEL PATIENT SIMILARITY NETWORK DRIVEN BY FEDERATED DATA QUALITY AND RESOURCE PROFILING

Alramzana Nujum Navaz

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Informatics and Computing

November 2023

United Arab Emirates University Doctorate Dissertation
2023: 66

Cover: Image regarding Patient Similarity Network Multidimensional Fusion Model with Data Quality Assessment and Resource Profiling in this study
(Photo: By Alramzana Nujum Navaz)

Print: University Print Service, UAEU 2023

Declaration of Original Work

I, Alramzana Nujum Navaz, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this dissertation, entitled *"A Novel Multi-Model Patient Similarity Network Driven By Federated Data Quality And Resource Profiling"*, hereby, solemnly declare that this dissertation is my own original research work that has been done and prepared by me under the supervision of Prof. Mohamed Adel Serhani, in the College of Information Technology at the UAEU. This work has not previously been presented or published, or formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my dissertation have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this dissertation.

Student's Signature: 

Date: 10.11.2023

Advisory Committee

1) Advisor: Dr. Mohamed Adel Serhani

Title: Professor

Department of Computer Science and Software Engineering

College of Information Technology

2) Co-Advisor: Dr. Hadeel El-Kassabi

Title: Assistant Professor

Faculty of Applied Sciences & Technology

Humber College Institute of Technology & Advanced Learning, Canada

3) Member: Dr. Mohammad Mehedy Masud

Title: Associate Professor

Department of Computer Science and Software Engineering

College of Information Technology

4) Member: Dr. Hany Alashwal

Title: Associate Professor

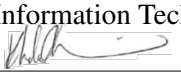
Department of Computer Science and Software Engineering

College of Information Technology

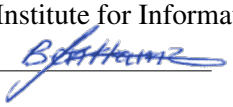
Approval of the Doctorate Dissertation

This Doctorate Dissertation is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Dr. Mohamed Adel Serhani
Title: Professor
Department of Computer Science and Software Engineering
College of Information Technology
Signature  Date 10.11.2023

2) Member: Dr. Khaled Shuaib
Title: Professor
Department of Information Systems and Security
College of Information Technology
Signature  Date 10.11.2023

3) Member: Dr. Fady Alnajjar
Title: Associate Professor
Department of Computer Science and Software Engineering
College of Information Technology
Signature  Date 10.11.2023

4) Member (External Examiner): Dr. Abdessamad Ben Hamza
Title: Professor
Concordia Institute for Information Systems Engineering, Canada
Signature  Date 10.11.2023

This Doctorate Dissertation is accepted by:

Acting Dean of the College of Information Technology: Dr. Fekri Kharbash

Signature  _____

Date 08/02/2024

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature 

Date 16/02/2024

Abstract

Smart and Connected Health (SCH) is revolutionizing healthcare by leveraging extensive healthcare data for precise, personalized medicine. At its core, SCH relies on the concept of patient similarity, which involves the comparative analysis of newly encountered patients with those who exhibit comparable similarities from the existing patient cohort. Yet, this approach faces significant challenges, including data heterogeneity and dimensionality. Our research introduces a multi-dimensional Patient Similarity Network (PSN) Fusion model tailored to handle both static and dynamic features. The static data analysis focuses on extracting contextual information using Bidirectional Encoder Representations from Transformers (BERT), while dynamic features are captured through neural networks and with Long-Short-Term Memory (LSTM) based autoencoders to reduce dimensionality while preserving temporal features. The key to our approach is the novel Similarity Network Fusion (SNF) scheme, that aggregates static and dynamic PSN similarity matrices. Compared to conventional classification methods, our deep learning-based PSN Fusion model demonstrates superior classification accuracy across various patient health outcomes. However, during our evaluation, we identified certain quality issues in the data that need to be addressed at each of the data value chain's processes to maximize the PSN's accuracy. Our Data Quality Management model introduces the data profiling concept to capture, enhance, and validate data quality at every stage of the PSN. We proposed Federated Data Quality Profiling (FDQP), inspired by Federated Learning, to extend the concept of quality profiling to the edge node, ensuring robust data quality assurance in distributed environments. It employs

federated feature selection and lightweight profile exchange, to swiftly identify and rectify data discrepancies. Extensive experiments across edge nodes demonstrate the positive impact of FDQP on data quality and the accuracy of Federated Patient Similarity Network (FPSN) models. Finally, we proposed a hybrid resource-aware FPSN solution to effectively combine static and dynamic resource quality profiles with edge reputation data to improve edge node selection. This all-inclusive approach ensures improvements in convergence time, as well as efficient memory, network, and disk usage within FPSN models. In summary, our research integrates the PSN Fusion model, the Federated Data Quality Profiling model, and the Resource-Aware Federated Profiling model to offer a holistic solution. This approach promises transformative benefits by efficiently managing multi-dimensional heterogeneous health data, emphasizing data quality, and optimizing edge computing resources. Ultimately, the goal is to deliver an enhanced healthcare system that empowers healthcare practitioners with efficient and data-driven tools, leading to improved patient outcomes.

Keywords: Smart and Connected Health, Precision Medicine, Patient Similarity Network (PSN), Data Quality Management, Edge Node Selection, Resource Optimization, PSN Fusion model, Federated Data Quality Profiling, Resource-Aware Federated Profiling, Federated Learning, Federated Patient Similarity Network (FPSN).

Title and Abstract (in Arabic)

شبكة تشابه المرضى متعددة النماذج مبتكرة تعتمد على جودة البيانات الموحدة وتوصيف الموارد

الملخص

تُحدث الصحة الذكية والمتصلة (SCH) ثورة في مجال الرعاية الصحية من خلال الاستفادة من بيانات الرعاية الصحية الشاملة للحصول على علاج دقيق وشخصي. تعتمد SCH في جوهرها على مفهوم تشابه المرضى، والذي يتضمن التحليل المقارن للمرضى الذين تم التعرف عليهم حديثًا مع أولئك الذين يظهرون تشابهات مماثلة من مجموعة المرضى الحالية. ومع ذلك، يواجه هذا النهج تحديات كبيرة، بما في ذلك عدم تجانس البيانات وأبعادها. يقدم بحثنا نموذجًا متعدد الأبعاد لشبكة تشابه المرضى (PSN) مصممًا للتعامل مع الميزات الثابتة والديناميكية. يركز تحليل البيانات الثابتة على استخراج المعلومات السياقية باستخدام تمثيلات التشفير ثنائية الاتجاه من المحولات (BERT)، في حين يتم التقاط الميزات الديناميكية من خلال الشبكات العصبية ومع أجهزة التشفير التلقائية المستندة إلى الذاكرة طويلة المدى (LSTM) لتقليل الأبعاد مع الحفاظ على الميزات الزمنية. إن مفتاح النهج الذي نتبعه هو مخطط دمج شبكة التشابه (SNF) الجديد، الذي يجمع مصفوفات تشابه PSN الثابتة والديناميكية. بالمقارنة مع طرق التصنيف التقليدية، يُظهر نموذج دمج شبكة تشابه المرضى (PSN Fusion) القائم على التعلم العميق دقة تصنيف فائقة عبر مختلف النتائج الصحية للمرضى. ومع ذلك، أثناء التقييم الذي أجريناه، حددنا بعض مشكلات الجودة في البيانات التي يجب معالجتها في كل عملية من عمليات سلسلة قيمة البيانات لتحقيق أقصى قدر من دقة PSN. يقدم نموذج إدارة جودة البيانات لدينا مفهوم ملفات تعريف البيانات لالتقاط جودة البيانات وتحسينها والتحقق من صحتها في كل مرحلة من مراحل PSN. لقد اقترحنا ملفات تعريف جودة البيانات الموحدة (FDQP)، المستوحاة من التعلم الموحد، لتوسيع مفهوم ملفات تعريف الجودة إلى محطات الحافة، مما يؤكد ضمان قوي لجودة البيانات في البيئات الموزعة. فهو يستخدم اختيارًا متحدًا للميزات وتبادلًا خفيفًا لملف التعريف، لتحديد وتصحيح تناقضات البيانات بسرعة. تُظهر التجارب المكثفة عبر محطات الحافة التأثير الإيجابي لـ FDQP على جودة البيانات ودقة نماذج شبكة تشابه المرضى الموحدة (FPSN). أخيرًا، اقترحنا حل FPSN مختلطًا مدركًا للموارد للجمع بشكل فعال بين ملفات تعريف جودة الموارد الثابتة والديناميكية وبيانات سمعة الحافة لتحسين اختيار محطات الحافة. يضمن هذا النهج الشامل إجراء تحسينات في وقت التقارب، بالإضافة إلى كفاءة استخدام الذاكرة والشبكة والقرص ضمن نماذج FPSN. في المجمل، يدمج بحثنا نموذج اندماج شبكة تشابه المرضى، ونموذج إدارة جودة البيانات الموزعة،

ونموذج الملف الشامل المدرك للموارد لتقديم حلاً شاملاً. هذا النهج يعد بفوائد محورية من خلال إدارة بيانات الصحة متعددة الأبعاد والمتنوعة، مع التركيز على جودة البيانات، وتحسين موارد الحوسبة على الحوسبة الطرفية. في نهاية المطاف، الهدف هو تقديم نظام صحي محسّن يمكّن ممارسي الرعاية الصحية بأدوات فعّالة مدعومة بالبيانات، مما يؤدي إلى تحسين نتائج المرضى.

مفاهيم البحث الرئيسية: الصحة الذكية والمتصلة، الطب الدقيق، شبكة تشابه المرضى، إدارة جودة البيانات، اختيار محطات الحافة، تحسين الموارد، نموذج دمج شبكة تشابه المرضى، ملفات تعريف جودة البيانات الموحدة، نموذج الملف الشامل المدرك للموارد، التعلم الموحد، شبكة تشابه المرضى الموحدة.

List of Publications

This dissertation is based on the work presented in the following papers, referred to by Roman numerals:

I. A. N. Navaz, M. A. Serhani, H. T. El Kassabi, N. Al-Qirim, and H. Ismail, "Trends, Technologies, and Key Challenges in Smart and Connected Healthcare," *IEEE Access*, vol. 9, pp. 74044-74067, 2021, doi: 10.1109/ACCESS.2021.3079217.

II. A. N. Navaz, H. T. El-Kassabi, M. A. Serhani, A. Oulhaj, and K. Khalil, "A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine," *Journal of Personalized Medicine*, vol. 12, no. 5, p. 768, May 2022, doi: 10.3390/jpm12050768.

III. A. N. Navaz, M. A. Serhani, H. T. El Kassabi, and I. Taleb, "Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling," *Sensors (Basel)*, vol. 23, no. 14, pp. 1–32, 2023, doi: 10.3390/s23146443.

IV. A. N. Navaz, H. T. El Kassabi, M. A. Serhani, and E. S. Barka, "Resource-aware Federated Hybrid Profiling for Edge Node Selection in Federated Patient Similarity Network," *MDPI Applied Sciences*, vol. 13, no. 23, p. 13114, 2023, doi: 10.3390/app132413114.

Author's Contribution

Alramzana Nujum Navaz's significant contributions to the dissertation have profoundly impacted various critical research aspects, including meticulous research, data compilation, literature surveys, and effective research strategy development.

I. Played a pivotal role in identifying and articulating the research objectives related to the multi-model PSN, contributing to the development of a comprehensive solution for its implementation.
II. During the research planning phase, assumed primary responsibility for tasks such as data collection, data processing, evaluation of results, and actively oversaw experimental work.
III. The comprehensive approach thoroughly addressed all dissertation objectives, demonstrating a deep understanding of the subject matter.
IV. Contributions extended to the conceptualization of essential ideas related to PSN, fusion algorithms, formal model, and architecture.
V. Led the dissertation project by formulating the research scope, identifying key exploration concepts, defining the chosen methodology, reporting research results, and providing a comprehensive background for enriched study conclusions.
VI. Exemplified leadership and sole responsibility in designing and executing the research experiments related to the multi-model PSN, instrumental in the dissertation's success.

Author Profile

Alramzana Nujum Navaz holds a Bachelor's degree in IT Engineering from Cochin University of Science and Technology, India, which she obtained in 2000. She furthered her academic journey by earning an MSc in IT Management from the College of IT at UAE University (UAEU) in 2017. Her primary research focus revolves around the application of intelligence in healthcare systems, with a profound aim to enhance patient care and save lives.

During the period from 2012 to 2018, she served as a Research Assistant at UAE University, where she actively contributed to various research endeavors. Her professional background spans over 15 years, encompassing extensive experience in application development, utilizing cutting-edge technologies such as deep learning, web services, and mobile APIs. Her passion lies in exploring the realms of smart eHealth, big data analytics, artificial intelligence, data mining, and mobile computing, which drive her continuous pursuit of innovation and excellence in the field.

Alramzana Nujum Navaz's commendable contributions to her field of study have been clearly demonstrated through her high-quality publications in top-tier journals and conferences. She is an accomplished teaching assistant at UAE University, and has successfully completed the PhD Teaching Academy Program conducted by The Center for Excellence in Teaching and Learning (CETL) The College of Graduate Studies (CGS) - UAEU. This dual expertise in research and teaching has not only enhanced her capabilities but also established her as a promising leader in the interdisciplinary fields of IT and healthcare.

Acknowledgements

I am grateful to Allah (S.W.T) for granting me the strength, intelligence, and abilities to complete this research. I extend my sincere appreciation to my supervisor, Dr. Mohamed Adel Serhani, and co-advisor, Dr. Hadeel El Kassabi, for their unwavering support, valuable insights, and extensive knowledge throughout my PhD journey. They have been exceptional mentors, guiding me every step of the way in my e-health research, and I am deeply thankful for their guidance. I'm grateful to my committee members, Dr. Masud and Dr. Hany, for their support and advice during my thesis work. My thanks to the CIT faculty, including Dr. Ezedin, Dr. Nazar, and Dr. Mamoun, for their guidance and inspiration. My co-authors Dr. Sujith, Dr. Ikbal, Dr. Oulhaj, Dr. Saad, Dr. Heba and K.Khalil, deserve my gratitude for their contributions. I acknowledge the vital roles played by Prof. Ali Hassan, Dr. Farag, Dr. Lakas, and Ms. Maryam Mandhari in supporting my PhD journey. I'm sincerely grateful to UAE University for the scholarship that supported my academic aspirations in IT. I'm deeply appreciative of my friends, including Tetiana, Balqis, Faiza, Zeid, Eiman, Nisha and Muhsina for their unwavering support. Special thanks to my dear friend Sabeena for instilling the "I can" mantra in me. I am deeply grateful to my parents, Safiya Beevi & E.M Basheer, and my mother-in-law, Rukiya Beevi, for their generous support, blessings and duas. My brothers Alsajir and Aljaseer, along with Sumina and Shahmi, deserve my thanks. I wholeheartedly appreciate the Kodikkakom & Nujum's Thottathil families for their love and guidance. I owe a debt of gratitude to my inspiring kids, Afzal and Aliya, and to my partner, Nujum Navaz, whose unwavering support and love made this dream achievable.

Dedication

To my wonderful parents, loving family and thoughtful friends

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

AI	Artificial Intelligence
DQ	Data Quality
FL	Federated Learning
MI	Multiple Imputations
ML	Machine Learning
SI	International System of Units
DQA	Data Quality Assessment
DQP	Data Quality Profile
DQA-FPSN	Data-quality-aware Federated PSN
EHR	Electronic Health Record
EWP	Edge Workload Profile
EWPM	Edge Workload Profiling Module
FDQ	Federated Data Quality
FHR	Fetal Heart Rate
FDQP	Federated Data Quality Profiling
FDQProfile	Federated Data Quality Profile
FPSN	Federated Patient Similarity Network
FWP	Federated Workload Profiling
IID	Independent and identically distributed
IoT	Internet of Things
LDQP	Local Data Quality Profiling
LDQProfile	Local Data Quality Profile
MCS	Mobile Crowd Sourcing
NRMA	Node Resource Monitoring Agent
NSM	Node Selection Module
PSN	Patient similarity network
SCH	Smart Connected Health
TDQM	Total Data Quality Methodology
XML	Extensible markup language

Chapter 1: Introduction and Background

Over the past decade, the healthcare industry has undergone a significant transformation, with the collection of massive amounts of patient data. This data is both complex and diverse, accounting for approximately 30% of the world's data volume, and is projected to grow at a compound annual rate of 36% by 2025 [1]. Using this data effectively will pave the way for new insights, therapies, drugs, and personalized care for several healthcare stakeholders. An intelligent health system makes it possible to provide individualized and precisely targeted care as the realms of physical and virtual medicine continue to merge [2].

The global digital health market (e.g., electronic health records (EHR), health monitoring, and wearable devices) was worth an estimated $175 billion in 2019, and it is projected to grow to nearly $660 billion by 2025 [3]. Additionally, the healthcare predictive analytics market was worth $7.88 billion in 2021, and experts predict that it will grow to $69.63 billion by 2029 [4]. This growth is driven by the increasing demand for personalized healthcare solutions, advancements in big data analytics, progression in EHR adoption, and the rising prevalence of chronic diseases [5].

The future of healthcare and patient experience is smart and will be shaped by smarter technology, smarter algorithms, and smarter care models. Smart Connected Health (SCH) is an innovative and technology-driven healthcare approach that integrates the power of connectivity and intelligence to revolutionize the medical industry. Leveraging its hyper-connected and

intelligent approach to healthcare, SCH is certain to play a significant role in this market growth.

To keep up with the rapid changes in the industry, the healthcare ecosystem must adapt to the rapid changes occurring in the industry by adopting new technologies, acquiring new technical skills, and embracing hyperconnectivity. SCH is hyper-connected, intelligent, patient-focused, and reliable. The goal of SCH is to encourage the creation and adoption of cutting-edge computational methods that can efficiently gather, link, analyze, and interpret data from an extensive range of sources, including EHRs. The fast growth of medical services based on artificial intelligence (AI), as well as new developments in the healthcare industry, is driven by the need for data management and precise decision-making. The SCH discipline will catalyze precision medicine by creating tools that aid in producing patient categorization, diagnosis, and prognosis, as well as guiding therapy and prevention.

Smart cities, driven by technologies like IoT, big data, AI, and drones, have evolved significantly in developed nations. Amidst evolving healthcare practices, digital technologies gained prominence, with 65% of EU healthcare providers adopting them to complement medical practices [6]. This shift towards virtual health highlights the importance of seamless data exchange and access to resources like national health databanks for improved pandemic monitoring [7]. Looking ahead, the post-pandemic era underscores the need for secure, intelligent access to healthcare, enabling swift diagnoses and tailored treatments [8].

This healthcare transformation is embodied by Health Care 4.0, centered around SCH [9]. The convergence of healthcare, IT, and mobile technology has led to digital tools capable of monitoring, analyzing, educating, and promoting well-being [10]. The Global Innovation Index 2022 [11] identifies two innovation waves: the Digital Age wave (supercomputing, AI, automation) and the Deep Science wave (biotech, nanotech, new materials). When applied to healthcare, these waves could revolutionize the industry, enhancing care delivery, patient experiences, clinician satisfaction, and outcomes.

In clinical settings, the use of low-quality digital information for diagnostic, therapeutic, and prognostic purposes introduces patient safety concerns, including misinterpretation of data, incorrect diagnoses, and treatment errors [12]. To tackle these issues, healthcare providers must prioritize elevating the precision and fidelity of digital health information. Furthermore, addressing data heterogeneity, and dimensionality requires implementing data preprocessing techniques like feature selection, dimensionality reduction, and data normalization. Although SCH has shown progress, there's a need for more research, dissemination, and impact to fully unlock its potential in advancing patient-centric care models. As the healthcare landscape undergoes a transformative shift, investing in SCH healthcare solutions becomes crucial. These solutions should strike a balance between data interoperability, data quality, performance accuracy, usability, resource utilization, and personalization.

To realize the vision of Health Care 4.0 [9] and effectively address the opportunities and challenges in this field, it's imperative to foster collaborative

endeavors involving governments, healthcare providers, researchers, policymakers, technology firms, and other stakeholders. At the outset, we conducted a systematic literature study on SCH and categorized state-of-the-art SCH technologies, defined SCH characteristics, identified enabling technology-related problems in SCH adoption, and proposed an architectural system in this research effort.

Researchers have developed a novel precision medicine approach known as the Patient Similarity Network (PSN), which leverages the existing SCH technologies. This approach facilitates the identification of the most similar patient, thereby providing clinicians with a valuable tool to gain insights from previously collected health data. Consequently, we conducted further research on PSN with the aim of facilitating the integration of health data from many sources into clinical practice. In light of the challenges faced by data heterogeneity and dimensionality, we proposed our multi-dimensional fusion strategy. However, we identified Data Quality (DQ) concerns that could arise at various stages of the PSN Data Value Chain, spanning from the initial source to the final stage, and we approached the problem holistically, covering all phases of the data value chain, resulting in our extended model. In addition, we examine the optimization and tuning of resources in order to enhance the performance of edge nodes utilized in Federated PSN (FPSN) with the objective of attaining greater accuracy. Thus, we envisioned a multi-model approach that combines and integrates health data from various sources and streamlines the process of data analytics along the Data value chain, as listed below.

1. Multidimensional PSN Data Fusion Model - Employs a novel fusion strategy to handle data heterogeneity and dimensionality, enabling comprehensive patient similarity analysis across diverse health data sources.
2. Data Quality Aware PSN Model - Aims to improve and ensure the quality of data at each stage of the data value chain, resulting in increased accuracy when forecasting similarities among patients.
3. PSN Resource Optimization Model - To optimize the resources through federated resource profiling to improve the overall performance of PSN that aids in edge node selection.

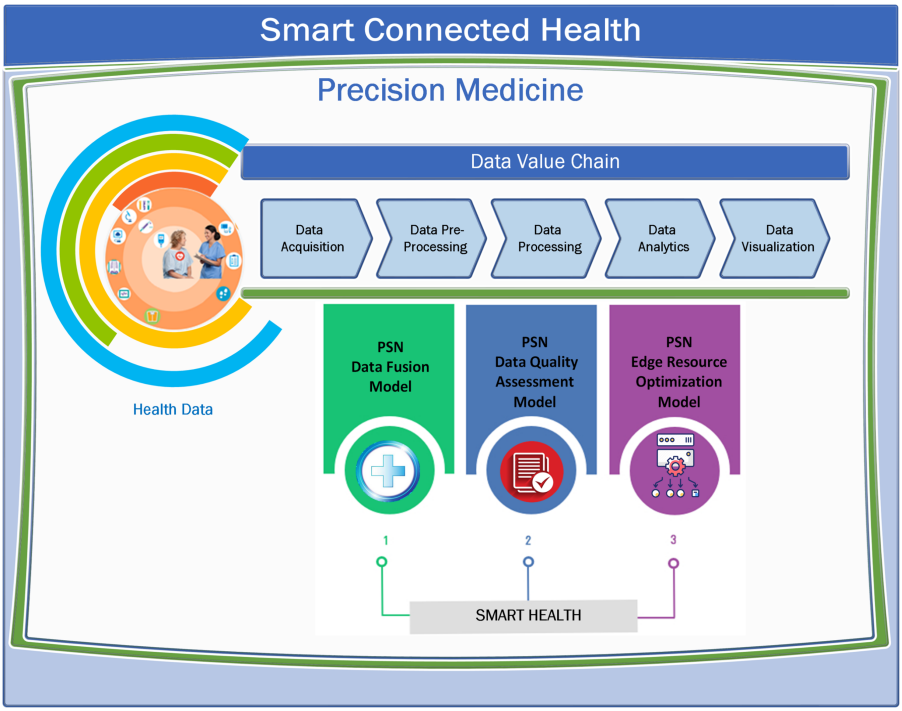


Figure 1: Multi-model PSN Approach to Enable Smart Health

The proposed PSN multi-model approach is conceptualized within the framework of SCH, under the umbrella of precision medicine. It emphasizes the healthcare data value chain, which involves the acquisition, preprocessing, processing, analysis, and visualization of health data from various sources, as shown in Figure 1.

1.1 Background

This section presents the main contexts of our study, including Smart Connected Health, Precision Medicine, and PSN, as well as the techniques of Deep Learning, Federated Learning, and their characteristics. In addition, we will emphasize the significance of data quality management in maintaining high-quality data within the healthcare data value chain.

1.1.1 Smart and Connected Health

Smart and Connected Health (SCH) encompasses interconnected digital healthcare solutions with remote capabilities [13]. In 2013, the National Science Foundation (NSF) and the National Institutes of Health (NIH) in the USA initiated the "Smart and Connected Health SCH: Connecting Data, People, and Systems" program. Its objective was to accelerate innovative information technology approaches in healthcare [14]. This program aimed to foster multidisciplinary research collaborations for pioneering "smart" healthcare solutions.

The integration of digital technology, including artificial intelligence, has seen significant progress in healthcare delivery. Deloitte groups telecare, telehealth, telemedicine, m-Health, digital health, and e-Health services under the umbrella of connected health or technology-enabled care (TEC) [15]. SCH

is an interdisciplinary research field at the intersection of medical informatics, public health, big data, bioengineering, and telecommunications. SCH's adaptability allows it to address diverse critical health contexts, facilitating resource-aware, time-constrained, complex, and secure healthcare transactions among various stakeholders.

SCH is revolutionizing healthcare with potential benefits such as faster treatment, lower physician visit costs, improved emergency response, and enhanced patient care [16]. Emerging technologies like AI, Big Data, and Cloud computing are reshaping health services. The widespread use of smartphones and tablets has led to a surge in mobile health apps and biosensing wearables, providing real-time clinical data access. SCH empowers both patients and healthcare practitioners, allowing greater control over well-being, online knowledge access, seamless communication with physicians, and access to support and interventions. During the COVID-19 pandemic, SCH has proven its efficiency in pandemic management, tracking, and risk mitigation strategies [17].

Inevitably, physicians will embrace the emerging world of technology, which is reshaping the practice of medicine. Although technology might initially cause hesitation, embracing SCH can advance precision medicine research and enable clinicians to make well-informed treatment decisions for their patients.

1.1.2 Precision Medicine

Precision medicine provides personalized treatment strategies for patients' subgroups, whereas traditional medicine uses a one-size-fits-all

approach to treat all types of patients. Currently, traditional medicine is shifting from a population-centric to a patient-centric strategy and can bring prevention, personalization, and precision to everyday medical practice. This shift enables the identification of hidden patterns within healthcare data. Precision medicine, an innovative and recent paradigm, considers individual differences in people's genes, environmental contexts, and lifestyles. Some variables may be more important to one subgroup than others and, certain treatment responses can differ across patients. This inspires doctors and medical researchers to devise novel techniques for detecting and analyzing subgroups [18].

In 2015, President Obama launched the Precision Medicine Initiative [19], which empowers individuals to take charge of their health by offering personalized healthcare solutions. In the UAE, the Emirati Genome Initiative has partnered with SEHA (Abu Dhabi Health Services Co.) to promote Emirati citizen engagement and employ genomic data for formulating healthcare policies that cater to present and future individual population needs, while also promoting the progress of preventive medicine [20].

Individuals often turn to peers in their field for guidance when making decisions about various life-related matters. Students may seek advice from seniors who have taken similar academic and career paths. Physicians may learn from their experiences handling diverse medical cases [21], and patients may seek recommendations and medical treatments from those who have encountered similar health conditions. Patient-friendly social platforms, such as PatientsLikeMe [22], are spaces where individuals with every type of condition share their health experiences, connect with similar patients, learn self-health

management, and engage in their well-being. These platforms facilitate patient-professional information exchange, ultimately enhancing patient care and accelerate realistic medical research initiated by patients.

Precision medicine's fundamental principle entails preventive and treatment techniques that factor in individual variations. This involves analyzing huge datasets that contain patient information, medical imaging, and genetic sequences [23]. The domain of biomedical and health informatics is rapidly expanding across all scales. SCH that includes medical devices, big data, cloud technologies, and personalized health advancements are enabling elements for the evolving landscape of precision medicine. This dynamic evolution underscores the necessity for novel scaling tools, integration frameworks, intelligent algorithms, and methodologies.

1.1.3 Patient Similarity Network

In the realm of precision medicine, Patient Similarity Network Networks (PSN) are a cutting-edge paradigm that groups patients based on shared characteristics, including genetic profiles, for more personalized and targeted healthcare interventions. PSN exhibits robust performance, boasts comprehensibility, and prioritizes patient confidentiality. The theory associated with the case similarity of patients can be explained using the following analogy: If two patients exhibit similarities across laboratory investigations, imaging, and clinical parameters such as age, genetics, and comorbidity status [24], their medical case progression is also likely to be similar [25]. Therefore, identifying past patients who are similar to the current patient can provide valuable insights into to disease investigations and potential treatments. Thus, the objective of PSNs is to recommend the appropriate therapy, medicine, and

lifestyle changes to the current patient based on relevant data extracted from similar patients, thereby outlining potential clinical outcomes [26]. Patient similarity applications can be extended to construct cohorts for cross-institutional observational research, disease surveillance, and clinical trial enrollment. Defining a patient similarity measure is thus a key step in allowing patients to be stratified into clinically meaningful subgroups [27].

1.1.4 Deep Learning

Deep learning (DL) has attracted exceptional interest in the recent decade, for its application in the study and diagnosis of biological and healthcare issues. The methodology has a history of uncovering relevant features and completing tasks that were previously difficult to tackle using alternative approaches and human experts. In the wake of big data, DL innovation is a growing trend for data representation and analysis. It is a form of machine learning approach that can decipher medical big data by cascading deeper hidden layers to form the DL network [28].

In estimating mortality risk among hospitalized patients, feedforward neural networks consistently outperform logistic regression and severity of illness scores. Recent advancements showcase the potential of neural architectures based on LSTM and other data mining techniques that can predict inpatient mortality, 30-day unplanned readmission, prolonged length-of-stay, and diagnoses using comprehensive EHR data [29].

Recently, novel DL architectures customized for survival analysis, a comparable time-to-event regression challenge with right censoring, have been introduced. Notably, as early as 2016, recurrent neural networks could

categorize dozens of acute care diagnoses in clinical time series of varying lengths. A number of researchers have conceptualized phenotyping as multi-label classification, with hidden layers of neural networks implicitly capturing comorbidity patterns [30].

1.1.5 Federated Learning

As businesses adopt the edge computing paradigm to accommodate the surging number of smart devices, Federated Learning (FL) and DL converge to create a powerful approach. This synergy not only distributes data processing to the network edge but also facilitates collaborative and privacy-preserving model training across decentralized devices. Hence, businesses have started shifting their data storage and computing closer to their networks' edge, giving rise to the edge computing paradigm.

Edge computing nodes have a finite amount of processing power, so effective resource allocation and management strategies are required to make the most of those capabilities and keep data processing times to a minimum. However, the high variability in the content requests makes prediction demand not trivial, and most classical prediction approaches require the gathering of personal information about users at a central unit, which raises privacy concerns for many of those users. In this setting, federated learning (FL) [31][32] has emerged as a promising strategy for safely executing learning procedures on data shared amongst a group of users. The key to FL is that it relies on sharing model parameters that can be aggregated to a joint model rather than sharing the data itself among non-trusting parties. FL follows a client-server architecture system [33], where a central server coordinates the training process, constructs the model, and makes it accessible to all participating clients.

Over recent years, significant research has been devoted to strengthening the FL domain, which has had an impact on performance metrics like accuracy and computational efficiency [34]. Ongoing endeavors aim to enhance the performance of FL systems by optimizing communication and computation resource utilization.

1.1.6 Data Quality Management

Recent advancements in SCH have benefitted social networks, sensor networks, and other internet-based applications in the healthcare domain. While it is apparent that data analysis can yield important insights, the findings of such an analysis are unlikely to be trustworthy unless the data is subjected to well-defined and effective verification and quality control processes prior to its utilization. Therefore, both academics and healthcare practitioners must prioritize the assessment of data quality when working with data. Data quality assessment is a requirement for data enhancement, aiming to determine the quality level of the data [35].

The data quality concept is essential for tracking data value and relevance, warranting assessment primarily throughout the pre-processing and processing stages of data transformations. Particularly with data derived from sources such as sensors, characterized by unstructured data lacking quality references, necessitates early-stage data profiling and assessment of specific quality dimensions. This underscores the need to continuously examine, enhance, and manage the quality of data attributes throughout its lifecycle, as it has a direct impact on the final insights obtained from data analysis. The viable strategy is an end-to-end DQ management scheme for improving data processing operations while reinforcing the quality of data. Profiling emerges

as an alternative approach to capture the quality overview, encompassing requirements, characteristics, dimensions, scores, and rules. However, guaranteeing DQ is a highly expensive and time-consuming procedure that necessitates a lot of computational resources [36].

1.1.7 Edge Computing and Resource Utilization

Edge computing describes a computational paradigm that brings data processing tasks closer to the source of data, or the "edge" of the network, instead of relying on centralized cloud locations [37]. This approach, predominantly seen in the IoT domain, aims to reduce latency, conserve bandwidth, and provide a swifter responsive computational service.

Resource utilization is intrinsically tied to the efficacy of edge computing. Given the constraints associated with edge devices, which could range from limited computational power to restricted energy capacities, optimizing resource utilization is of paramount importance. This optimization not only extends the lifespan of devices but also guarantees seamless and efficient functioning of edge applications [38].

In light of this, the Quality of Edge Computing Services becomes a crucial metric, that provides an assessment of the overall efficiency, reliability, and performance of the services rendered at the edge. The more adeptly resources are utilized, the superior is the quality of edge services offered to the end-users [39].

This background establishes a cohesive foundation, intertwining various technological advances essential to modern healthcare research. SCH represents the forefront of technological integration into healthcare, enabling a

platform for the emergence of Precision Medicine. This facilitates a more advanced, individual-focused approach, further refined through the PSN, which clusters patients based on shared attributes. DL and FL emerge as pivotal in navigating and analyzing the extensive data involved, ensuring robust, privacy-preserving, and insightful outcomes. Amidst this technological surge, edge computing emerges as a key player for efficient data processing, optimizing resource utilization and enhancing the Quality of Edge Computing Services. Within these interconnected domains, Data Quality Management is essential, ensuring the reliability and accuracy of the insights derived, which are vital for the practical application and integrity of healthcare research.

1.2 Motivation

Prior to the advent of technology, physicians relied on experience-based methods to forecast healthcare treatments, often by drawing similarities with past cases. However, as healthcare data volumes have grown substantially, the process of manually detecting patient similarity has become complex and time-consuming. Consequently, clinicians rely on algorithms that estimate patient similarities automatically, based on health and event-driven data from numerous patients. These strategies have a significant influence on improving patient outcomes in practice and enable physicians to prescribe medications and recommend lifestyle adjustments. Precision medicine, encompassing PSN, plays a pivotal role in achieving these goals.

Although PSN approaches are relatively recent in the realm of precision medicine, substantial challenges need addressing before the full promise of PSN can be realized. The healthcare data value chain is complex and has various points of inception and data integration accounting for the

entire patient journey including physician input. Furthermore, concerns related to the variety, volume, and veracity of big health data are of prime importance.

Developing an accurate patient similarity measure confronts challenges in capturing the medical events of patients without losing information. While some efforts have been made to apply patient similarities in different applications, significant barriers remain in systematically adopting effective patient similarity. Dealing with the high dimensionality and sparsity of patient's data poses a major challenge. Heterogeneous clinical narrative data have hidden information that is valuable in PSNs.

EHRs encompass a broad spectrum of healthcare data, ranging from diagnoses and medications to laboratory tests and various medical events, often presented in a complex, high-dimensional format. The medical events are temporally sensitive, and the temporal detail is critical for comprehending the dynamics of medical terminologies and inferences. The interpretation of temporal representation is extremely difficult when using noisy clinical datasets, and the accuracy of outcome prediction is low. One of the existing methods to integrate multiple biological data is to concatenate standardized measurements. However, the concatenation of data tends to dilute the data quality with noise. Existing patient representation models are generally inadequate in terms of clinical interpretations due to the complexity of medical data, which if addressed would considerably broaden their application. To overcome these issues, we present our multidimensional PSN data fusion model.

The proposed PSN fusion model demonstrates enhanced classification accuracy in identifying diverse patient health outcomes compared to conventional medical classification algorithms. However, we have identified DQ issues that may arise in the PSN Data Value Chain from the source to visualization and may affect the ultimate model's accuracy. Addressing DQ holistically across all phases of the data value chain is imperative, necessitating sophisticated methodologies like data quality profiling to sustain the benefits of PSN. The PSN model also faces challenges arising from underutilization of existing medical data due to siloed storage, as well as concerns regarding data transportation and privacy of health data.

Using PSN distance estimations from both static and real-time data, the patient similarity network fusion method highlights the similarity between patient pairs while reducing the influence of non-similar pairings. However, the quality of data received from the edge nodes, or the sources were not taken into account when designing this PSN fusion strategy.Our experiments measuring the effect of edge-level data quality on FL model accuracy prompted further exploration into data quality-aware edge selection and profiling for PSN. This integration with FL services addresses inaccurate data from multiple client locations.

In order to reduce the computational burden of FL training with numerous datasets and jobs, we propose a model to efficiently enhance data quality in edge clients. Through dynamic context-based profiles sent to FL clients, we guide data selection and augmentation on the client side, significantly reducing data transmission related to individual patients. Our motivation to develop a quality-driven edge-based federated strategy for

sensor-based monitoring setup led us to propose the DQA FDQ profiling model.

Leveraging the resource capabilities of edge computing augments the reliability and decreases response times of real-time paradigms heavily reliant on edge computing. Such paradigms include Federated Patient Similarity Network (FPSN) models that distribute processing at each edge node and fuse the built PSN matrices in the cloud, presenting a unique challenge in terms of optimizing training and inference times, while ensuring efficient and timely updates at the edge nodes. In response, we propose a resource-aware federated hybrid profiling approach, that measures both the static and dynamic resources available at the edge nodes.

To sum up, the healthcare data value chain is multi-faceted, and decisions taken based on data can have far-reaching implications. In PSN, each component of the data value chain should provide value to the original inputs and outputs, by augmenting data quality. This necessitates holistic consideration of each individual data points, as well as a broader perspective in the context of cross-cutting SCH. The goal of this research is to develop advanced smart data models and techniques, with a specific focus on Patient Similarity Networks (PSNs), to enhance the healthcare Data Value Chain, encompassing data quality, patient similarity analysis, and resource optimization. This aims to improve the precision and efficacy of healthcare forecasts and decision-making processes. Thus, the motivation of this work is to transform the existing healthcare value chain into a smart data-driven, process-driven, resource-optimized, and patient-centric, data value chain that

can deliver real-time insights, information, and decision support across operational and clinical systems.

1.3 Ethical and Privacy Considerations

In researching healthcare data, ethical considerations are vital. Using patient records, carries inherent risks such as potential privacy breaches and unintended disclosure of patient specifics. To address these concerns, our study strictly followed data anonymization protocols, ensuring that all patient data was free from personally identifiable information, making individual identities indiscernible from the datasets. Our choice to harness datasets like the Covid-19 Epidemiological Data and the Framingham Offspring Heart Study stemmed from our confidence in these databases' rigorous data collection and dissemination procedures, emphasizing patient privacy.

In furthering our commitment to privacy, the principles of FL, central to our FPSN and FDQP, prioritize the security of patient data. By advocating for data localization, where information stays at its source, these techniques significantly reduces potential data breaches during transfers.

Although many research endeavors emphasize continuous consent and oversight, our focus was on utilizing pre-existing, anonymized datasets. Our aim combined technological advancement in healthcare with an unwavering dedication to upholding the rights, privacy, and dignity of all represented individuals.

1.4 Problem Statement and Objectives

Our main objective in this research is to develop a novel solution to enhance the healthcare Data Value Chain by employing smart data models. Our

proposed initial model, the Multi-dimensional PSN Data Fusion Model utilizes Similarity Network Fusion (SNF) approach to manage heterogeneous and high-dimensional health data. The model is extended to account for data quality throughout the data value chain, and integrates Edge Resource optimization to fine-tune resource utilization within the context of FPSN.

The key research objectives of our proposed study are summarized below, while a more detailed breakdown, including our contributions, is depicted in Figure 2.

1. To comprehensively assess, classify, and consolidate existing SCH models and solutions, culminating in the creation of an expert-based classification model that provides valuable insights into the landscape of SCH methodologies, facilitating the design of an innovative and efficient patient-centric healthcare system.
2. To leverage the Multi-dimensional PSN Data Fusion Model, capable of integrating and analyzing both temporal and clinical narrative data, to significantly improve the accuracy of patient similarity analysis, thereby enhancing the effectiveness of precision medicine applications.
3. To establish an end-to-end data quality improvement strategy across the PSN data value chain, integrating data quality assessment methodologies to ensure the reliability and accuracy of healthcare insights.
4. To optimize the utilization of edge resources in Federated Patient Similarity Network (FPSN) models by introducing a resource-aware federated hybrid profiling approach, enhancing system performance, reducing latency, and guaranteeing Quality of Edge Computing Services for healthcare applications.

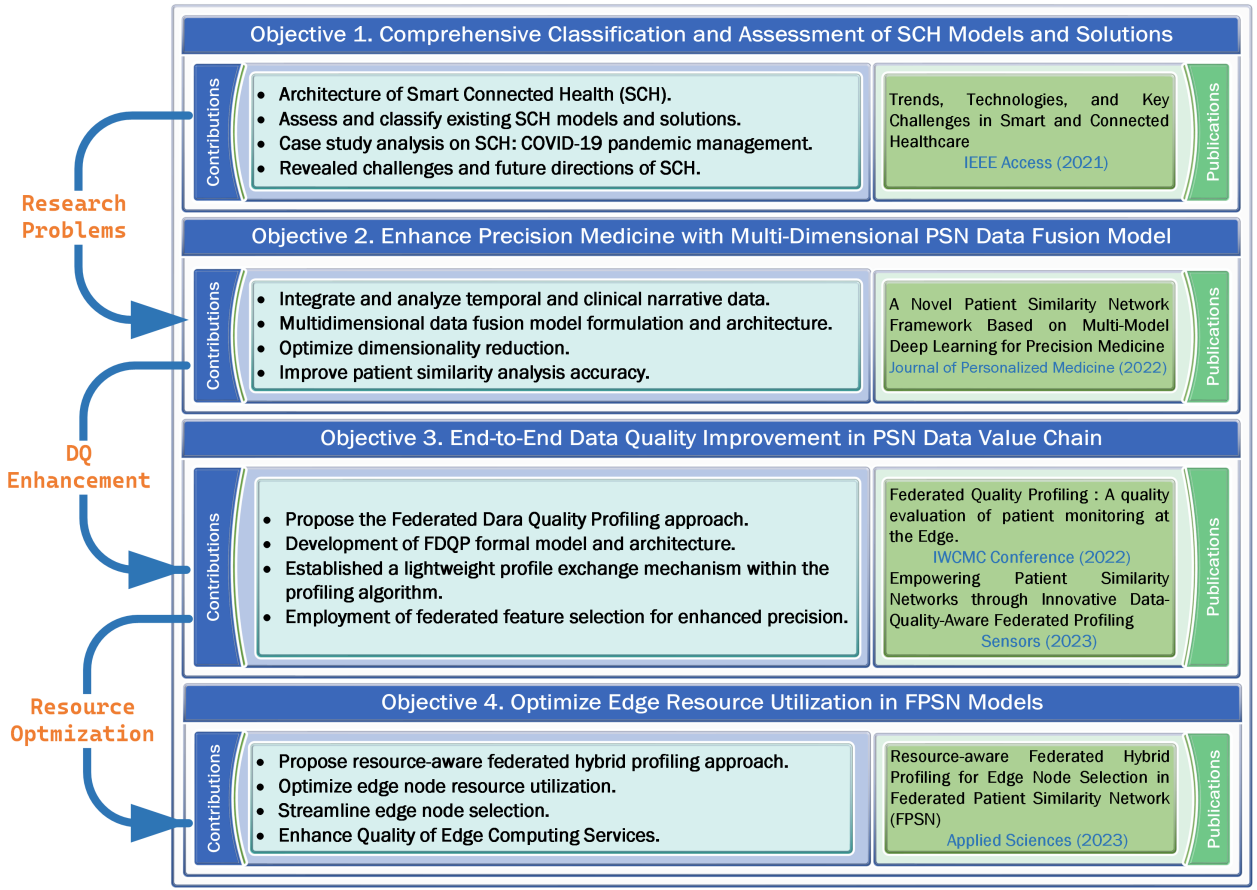


Figure 2: Research Objectives and Contributions

This research aims to answer the following research questions:

1.4.1 Research Question 1

How can we develop an effective SCH framework and strategies to provide smart, efficient, proactive, patient-centric healthcare? To answer this question, we propose the following contributions.

1. Devise an expert-based classification model and literature review on various existing SCH models, solutions, and architectures.
2. Propose an architectural model that captures the SCH solution's technical aspects, its environment, and the key stakeholders involved. Furthermore, the proposed model incorporates core technologies such as IoT, Big Data analytics, and AI to assist the primary stakeholders in enhancing healthcare efficiency.
3. A comprehensive case study is presented, outlining how the incorporation of SCH in conjunction with a range of technological advancements has served as a critical foundation for effectively managing and mitigating the repercussions of the COVID-19 pandemic.
4. Identify key SCH challenges and offer prospective research and technology paths for next-generation SCH, as well as provide recommendations for future SCH system implementations.

1.4.2 Research Question 2

How can we deal with heterogeneous health data including temporal information? How can high dimensionality in health data be dealt with and how to integrate data?

1. Our proposed approach addresses both temporal and clinical narrative data by implementing a hybrid model that considers the static and dynamic aspects of patient data in patient similarity analysis that will improve the accuracy. Static data modeling handles static patient profile data, whereas dynamic data modeling handles longitudinal dynamic data, where each patient is associated with a sequence of visits. Our incorporation of advanced Natural Language Processing (NLP) models like BERT facilitates the understanding of unstructured textual features.
2. Using the generalized hybrid model, the heterogeneity of the eHealth data from various data sources can be managed. As a result, the proposed model is efficient in addressing the big data challenges, where the structured and unstructured data of patient cases are diverse. Moreover, the reduction of dimensionality is a strategy developed within our model using autoencoder to achieve a robust and statistically sound machine learning model.
3. An innovation of significant importance lies in the patient Similarity Network Fusion (SNF) approach, that fuses the PSN distance calculations from static and dynamic data. This fusion not only emphasizes the similarity of the patient pair but also effectively mitigates the interference caused by non-similar pairs, thereby elevating the precision and reliability of analytics.

1.4.3 Research Question 3

Does enhancing data quality across the PSN data value chain improve its effectiveness and accuracy?

We believe that Data Quality assessment and improvement can further enhance the PSN classification Model accuracy and performance with our proposed PSN DQ Assessment Model, by achieving the following objectives.

1. Automate critical extract, transform, load (ETL) processes to always guarantee that data is of high quality. Devising a strategy for capturing Quality at the early data collection stages of the PSN model, to ensure data quality at the source.
2. Data quality management with Data Quality Profiling (DQP) with the creation and customization of data quality metrics according to the context and domain-specific characters of the data. Assessing data quality using data profiling can lead to some interesting use cases.
3. Continuous Adaptive data quality and assessment, monitor data pipeline, and ensure timely notifications regarding potential data corruption, thereby maintaining end-to-end reliability. This continuous assessment not only safeguards the data flow but also assesses pre-data quality and post-data quality impact on PSN computations. Such evaluations strengthen accuracy by addressing data errors and gaps.
4. Big Health data transportation and Data Privacy to be tackled by Data-driven federated efforts. This strategic handling enables effective data utilization without compromising confidentiality.

1.4.4 Research Question 4

How can we optimize the edge resources, employed in the FPSN? Can edge workload profiling aid in edge node selection and thus boost the overall performance of the FPSN?

This study presents a novel federated hybrid resource profiling method to assess the static and dynamic capabilities of edge nodes in edge computing systems. By efficiently allocating tasks based on available resources, the proposed approach enhances system performance, reduces latency, and guarantees Quality of Edge Computing Services. Leveraging federated workload profiling, and collaborative efforts between edge nodes further improves resource utilization, making it a comprehensive solution for edge computing optimization.

1. Establishment of an appropriate classification of edge resources, including static and dynamic attributes, to facilitate efficient resource utilization in edge computing.
2. Introduction of a resource-aware federated hybrid profiling strategy within a distributed edge computing environment.
3. Proposal of a reputation-based node selection approach for improved edge computing resource allocation.
4. Building the Edge Workload Profile (EWP) by continuously monitoring dynamic resources of edge workloads.
5. Development of the Federated Workload Profiling (FWP) model, incorporating federated resource profiling inspired by Federated Learning.
6. Utilizing edge performance metrics such as data training accuracy and convergence time to assess the efficacy of the proposed model in meeting the demands of the FPSN (Federated Patient Similarity Network).

Chapter 2: Literature Review

Healthcare innovations like telehealth, ehealth, mhealth, precision medicine, personalized medicine, and predictive health analytics are very appealing fields of study to researchers in academia and industry because they improve the quality of care for patients. Smart healthcare employs technologies, such as IoT, big data, cloud computing, and artificial intelligence, to enhance healthcare quality, increase operational efficiencies, and spur biomedical research breakthroughs in the treatment of disease. In this context, we first review the current state-of-the-art in SCH, and then classify them based on supporting technologies, application context, and futuristic systems. We further review the literature to see how PSN might be utilized as a classification tool to provide predictions to facilitate precision medicine. Then, we analyze the literature on data quality to get insights from the definitions of DQ dimensions, identifying DQ challenges, and the literature-based recommendations with few technical specifics on DQ enhancement. Further, we review the resource awareness in a federated edge setting. Finally, we summarize the knowledge gaps exposed by literature analysis.

2.1 Smart and Connected Health

In this subsection, we performed a scoping review of the literature and categorized the state-of-the-art on SCH as shown in Figure 3. The first three clusters looked at current SCH trends, advancements, and deployments in conjunction with AI, IoT, and other supporting technologies. Fourth and final clusters were added to focus on the application context of SCH and future SCH research.

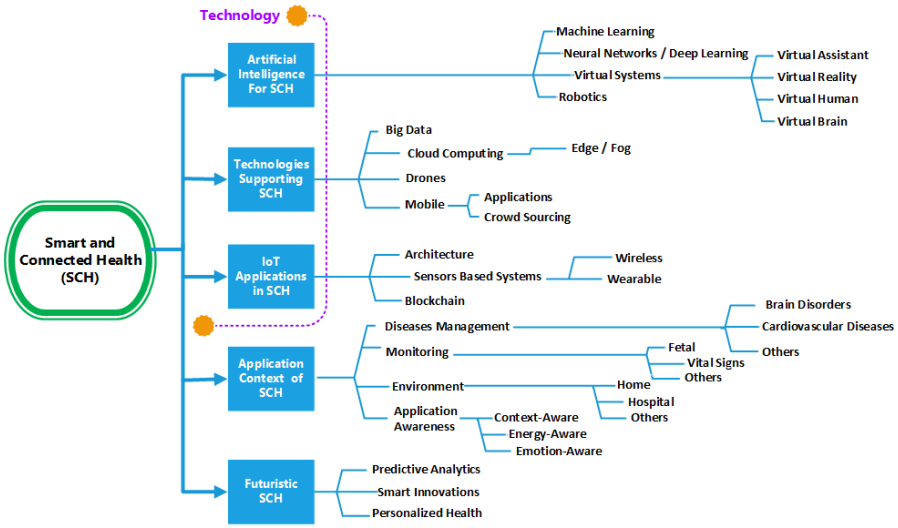


Figure 3: Classification of Smart and Connected Health

2.1.1 Artificial Intelligence for SCH

This cluster explores the healthcare impacts of AI and its various branches namely ML, DL, robotics, and virtual systems and the relevant studies are shown in Table 1. Similarities can be observed across several of these settings, while the distinctions are also readily apparent.

Table 1: Studies on AI-enabled SCH.

	AI-Enabled Smart and Connected Health						
	Machine Learning	**Deep Learning**	**Robotics**	**Virtual Systems**			
				Virtual Assistant	**Virtual Reality**	**Virtual Human**	**Virtual Brain**
Referenced Papers	[40], [41], [54]	[42], [43], [55], [56], [65], [66]	[44], [45], [46], [57], [58], [59], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81]	[47], [48], [60]	[49], [50], [61], [62]	[51], [48], [63], [64]	[52], [53]

2.1.2 *Technologies Supporting SCH*

This section investigates the data and processing technologies and platforms that support SCH, with the related studies shown in Table 2.

Table 2: Technologies Supporting SCH: Referenced Studies.

	Technologies Supporting SCH					
	Big Data	**Cloud Computing**		**Drones**	**Mobile Computing**	
		Fog	**Edge**		**Apps**	**Crowdsourcing**
Referenced Papers	[76], [82], [83], [95], [96], [97], [100], [101], [102]	[84], [85]	[86]	[87], [88], [98], [99]	[89], [56], [90]	[91], [92], [93], [94]

2.1.3 *IoT Applications in SCH*

The fourth industrial revolution in recent years has contributed to the so-called growth of the Internet of Robotic Things (IoRT) [57], integrating robotics, cloud computing, and IoT. IoT has become a pivotal technology in SCH, drawing substantial research interest [103]. Particularly within the realm of e-health, IoT offers significant advantages such as cost reduction, improved interoperability, user-friendly interfaces, real-time analytics, extended availability, and continuous monitoring [104].

Table 3: IoT Applications in SCH: Referenced Studies.

	IoT Applications in SCH			
	Architecture / Framework	**Sensor-based Systems**		**Blockchain**
		Wireless	**Wearable**	
Referenced Papers	[105], [104], [113], [114], [115], [121], [122], [126], [127], [128]	[106], [107], [116]	[108], [109], [117]	[110], [111], [112], [118], [119], [120], [123], [124], [125]

Research on IoT applications in SCH can be broadly categorized into three main areas: IoT architecture/framework, sensor-based systems, and IoT applications in blockchain, as summarized in Table 3.

2.1.4 Application Context of SCH

Studies referencing application context in SCH fall into four main categories: disease management, monitoring, environment, and application awareness as depicted in Table 4.

Table 4: SCH Application Context: Referenced Studies.

	Application Context of SCH			
	Disease Management	**Monitoring**	**Environment**	**Application Awareness**
Referenced Papers	[129], [130], [131], [109], [137], [138], [147], [51], [148], [41], [153], [154], [50], [133], [159], [139], [107], [56], [46], [40]	[132], [133], [120], [139], [140], [141], [49], [108], [126], [155], [156], [90], [117], [160], [161]	[44], [134], [135], [62], [142], [143], [139], [149], [150], [157], [61], [158], [153], [129], [134], [47], [148], [154], [109], [147], [148], [162], [154], [72], [60], [141], [133], [132], [140]	[86], [42], [136] [144], [145], [146], [151], [152], [101]

SCH is pivotal in medicine, serving roles from remote diagnosis to self-monitoring and disease management. Health monitoring covers diverse areas, including fetal heart rate monitoring [155] during pregnancy and elderly monitoring [49, 141] for conditions like Parkinson's or Alzheimer's disease.

2.1.5 Futuristic SCH

The future of SCH relies on smart hospitals and innovative solutions that enable advanced patient-physician interactions. In case of accidents, mobile apps can establish geographical connections with the nearest hospital

and transmit data before admission, potentially saving lives [143]. Anticipated developments include digitally enhanced self-care technologies, such as continuous glucose-monitoring systems where sensors serve as displays and wirelessly transmit data to emergency systems or mobile apps. Eye-tracking can guide the development of patient-centric drug delivery systems [137]. Predictive analytics will play a pivotal role in precise disease prediction and diagnosis, with the healthcare sector reaping more benefits than potential challenges [163]. Refer to Table 5 for referenced studies on SCH's futuristic aspects.

Table 5: Futuristic SCH: Referenced Studies.

	Futuristic SCH		
	Predictive Analytics	**Smart Innovations**	**Personalized Health**
Referenced Papers	[114], [115], [164], [165], [166], [167]	[106], [168], [169], [170]	[97], [171]

The current SCH systems leverage new technologies to create efficient, cost-effective, and fully connected monitoring solutions. Looking ahead, SCH is expected to evolve, offering flexible, cost-conscious solutions and services to benefit patients, healthcare practitioners, organizations, and governments. This will pave the way for next-generation personalized healthcare.

2.2 Patient Similarity Network

There have been several kinds of research on patient similarity as a core concern. As a clinical research technique, PSN is theoretically comparable to clinical diagnosis, which frequently includes a physician comparing a patient to a fundamental database of similar patients seen by them. In this system,

each input patient data feature is transformed into a patient similarity network (PSN) [172], where each PSN node represents an individual patient, and edges between patients reflect their pairwise similarity. The DL-based supervised patient similarity approach, as shown in Figure 4, represents patient pairs using embedding matrices (E_a and E_b). These matrices undergo convolutional filtering and map to feature maps for neural network training. This process generates deep embeddings (P_a and P_b) by pooling patient feature maps into intermediate vectors. A symmetrical similarity matrix with feature vectors calculates patient 'a' and 'b' similarity.

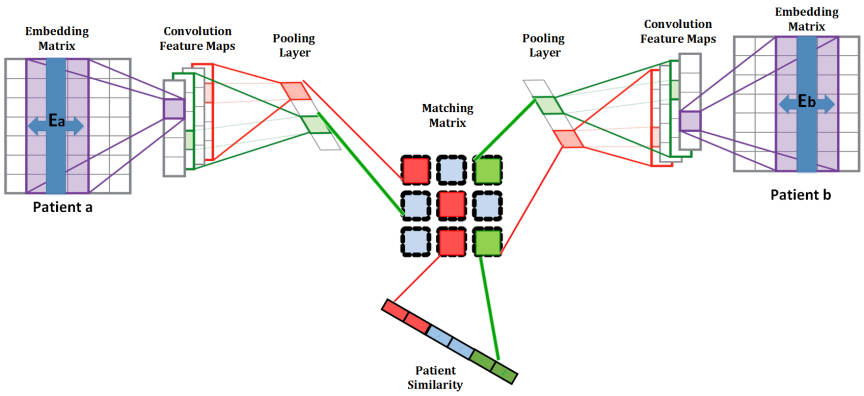


Figure 4: Supervised PSN Framework

In this section, we review the current literature on PSNs, focusing on approaches for building the network, combination models, healthcare application domains, and performance evaluation methods.

2.2.1 Distance Measurements in PSN

Distance measures play a fundamental role in patient similarity networks, impacting the performance of machine learning models [173]. In

this section, we discuss four commonly used distance measures: Euclidean, Manhattan, Cosine, and Jaccard, as summarized in the Table 6.

Table 6: Commonly Used Distance Measures in Patient Similarity Networks

Distance Measure	**Description**
Euclidean Distance	Measures the shortest distance between two points using Cartesian coordinates, suitable for low-dimensional data [174].
Manhattan Distance	Calculates distance between real-valued vectors without diagonal movement, often producing larger estimates than Euclidean distance [175].
Cosine Similarity	Computes the cosine of the angle between vectors, favored for high-dimensional data to overcome Euclidean distance issues [176].
Jaccard Index	Evaluates set similarity/diversity by comparing the number of shared entities to the size of their unions, often used for binary or binarized data [177].

These distance measures serve various purposes, from analyzing text document clustering to assessing similarity in healthcare data.

2.2.2 Existing Techniques for Building PSNs

Five categories [27] for building patient similarity are recognized in literature: clustering, dimensionality reduction, similarity, software tools, and a combination of clustering or similarity metrics and supervised approaches.

Table 7 presents a few of the approaches that exist for building patient similarity, including neural networks and DL.

In [178], two patient similarity measurement approaches were employed: one involved clustering patients using K-means and hierarchical clustering, while the other used a supervised technique with medication plans as class variables.

Table 7: Methods Used for Building Patient Similarity

Method	Parameters/Factors	Applications
Deep learning	ICD9	Unsupervised/supervised patient similarity (CNN) [179] Diagnosis with LSTM recurrent neural networks [180] Personalized disease prediction with (CNN) [181]
Triplet-loss metric learning	longitudinal EHRs	Personalized prediction [182]
Temporal similarity	Temporal sequences	Clinical (workflow) cases similarity [183]
Clustering	Variety of components of patient data	Patient similarity analytics loop [24]
Similarity measure construction	ICD code, Empirical co-occurrence frequency, Medical history, Blood test, ECG, Age, Gender	Predict individual discharge diagnoses [184] Predict ICU mortality [185]
Deep patient representation (three-layer stacked denoising autoencoders)	ICD9	Future disease prediction [186]
Similarity network fusion (SNF)	Nodes represent patients, and patients' pairwise similarities are represented by edges	Network-based survival risk prediction Identifying cancer subtypes [187]
Locally supervised metric learning (LSML)	Longitudinal patient data	Personalized predictive models and generation of personalized risk factor profiles [188]
Collaborative filtering methodology	ICD data	Creates a personalized disease risk profile and a disease management plan for the patient [189]
Anonymous indexing of health conditions for a similarity measure	Text similarity	Recommend two other patients foreach patient based on a keyword [190]
SimSVM	14 similarity measures from relevant clinical and imaging data	Predicting the survival of patients suffering from hepatocellular carcinoma (HCC) [191]
Concept hierarchy	Hierarchical distance measure	Detecting correlations in medical records by comparing the hierarchy of terms considering the distance between non-similar records in a hierarchy [192]

A variation of the influence-diagram representation, known as a similarity network [193], was introduced in the late 1990s for constructing complex influence diagrams. It consists of a similarity graph and local knowledge maps, where nodes represent hypotheses and edges connect similar hypotheses. Similarity networks extend the belief network representation, which was the foundation of Pathfinder [194], a decision-theoretic expert system for hematopathology diagnosis.

User Intelligent self-learning electronic medical record (ISLEMR) [195] is based on PSN and considers the principal diagnosis as the similarity assessment input. It can provide treatment plan recommendations and can help in training inexperienced doctors. Patients P_i and P_j had principal diagnoses D_i and D_j, respectively, and the similarity is 1 when D_i is the same as D_j and when they differ, the similarity is 0. In ISLEMR [195], physicians can choose from multiple learning modes when utilizing the self-learning feature. These modes include an "EASY MODE," where patients with the same principal diagnosis are grouped as the most similar, a "MIXED MODE" that considers demographic, vital signs, and lab test data, and a "COMPLEX MODE" that iteratively clusters objects based on similarity until all objects are allocated into two clusters. These modes offer flexibility in tailoring the system's recommendations and learning approach to specific patient cases and needs.

In [185], Lee et.al proposed a cosine-similarity-based patient similarity metric (PSM) to an intensive care unit (ICU) database to identify similar patients and then build 30-day mortality prediction models. The PSM is based on the cosine of the angle between two patient vectors called cosine similarity [185]. Two patient vectors in opposite directions would result in a value of

-1 (minimum similarity), whereas two identical patient vectors will result in a value of 1 (maximum similarity). In [183], the authors use the clinical similarity of workflow to discover cases by comparing the optimal case with specific patient instances, utilizing an interval similarity function based on the intra-task distance, which is the distance between intervals representing corresponding tasks, and the inter-task distance, which is the distance between the relations of corresponding tasks considering possible dissimilarities of cases in tasks occurring in sequence.

The International Classification of Diseases-Version 10 (ICD-10) encodes the semantic information of disease conditions, severity, diagnoses, and treatments. In the context discussed in [172], ICD-10 is utilized for supervised patient distance measurement, enabling patient clustering based on their ICD-10 associations. Patients are represented as vectors within a patient-ICD10 association matrix spanning 674 ICD-10 codes. To gauge the significance of ICD-10 occurrences, [196] employed the term frequency-inverse document frequency (TF-IDF) measure, utilizing cosine similarity to calculate the angles between vector pairs. This approach, as also seen in research on phenotype similarity [197], significantly enhances the quality of predictive data.

As per AMIA Symposium 2019 recommendations [198], patient similarities fall into four classes: feature, outcome, exposure, and mixed classes, distinguishing entity-based (e.g., tumors) from event-based (procedures) characteristics. Many patient similarity network (PSN) methods use vector-based patient representations, which aggregate medical procedures over time, potentially losing temporal information. A study by Zhu et al. [179]

proposed a Convolutional Neural Network (CNN) to evaluate patient similarity based on temporal alignment in longitudinal Electronic Health Records (EHRs) using medical concept embedding. This approach calculates patient similarity from the temporal representation, where each patient's records are structured as a matrix X, with dimensions d × Np, where d represents the dimension and Np indicates the total number of hospital visits for patient p.

Personalized predictive modeling involves patient similarity computation, feature filtering, predictive modeling, and risk factor profiling [188]. It employs a trainable similarity measure called locally supervised metric learning (LSML) to find patient similarity. The study suggests that static similarity measures like Euclidean or Mahalanobis distances may not be optimal for all disease conditions and proposes a logistic regression (LR) predictive model to compute risk factor profiles, particularly assessing the risk of diabetes onset.

Within a NoSQL database management system, PSNs are discussed, covering in-database data analysis, pre-processing, and patient similarity calculations [199]. In contrast, DeepPatient introduces an approach that generates more compact, lower-dimensional representations of EHRs using stacked denoising autoencoders, facilitating improved data scalability [186].

Concerning drug and gene-gene similarity, SITAR is an algorithm for predicting drug targets based on drug-drug and gene-gene similarity computations, featuring feature selection and prediction using logistic regression [200]. Semantic similarity metrics are employed to measure

phenotypic similarity based on the human phenotype ontology, ranking diseases [201].

In a federated framework for PSNs across organizations, a privacy-preserving platform is proposed to find similar patients from multiple hospitals without sharing patient-level information [202].

Precision epidemiology is advocated, considering factors such as the viral genome, host genome, disease course, and transmission history for a more comprehensive approach [203]. Additionally, a disease transmission dynamics map is introduced, focusing on the similarities and dissimilarities of disease dynamics among various countries [204].

2.2.3 Combination Neural Network Models

Various neural network (NN) combinations have been proposed in the literature, addressing diverse challenges such as patient-trial matching, protein classification, clinical semantic textual similarity, and data fusion [205–208]. Table 8 summarizes these NN combination models.

Table 8: NN Model Combinations for Various Healthcare and Data Challenges

Model	Purpose	Components	Performance
COMPOSE [205]	Patient-trial matching	BERT, CNN	98% accuracy
DeepPPPred [206]	Protein classification	RNN, CNN, BERT	Outperforms individual networks
Gated Network [207]	Clinical semantic textual similarity	BERT, One-hot	High Pearson correlation (0.8525)
Matrix Factorization [208]	Data fusion	Various data sources	Valuable in gene function prediction

2.2.4 *PSN Application in Various Health Domains*

Patient similarity research encompasses various medical domains, spanning diseases like glioblastoma multiforme (GBM), hepatocellular carcinoma, diabetes, schizophrenia, and various cancers. This field is also evolving to incorporate genomics, proteomics, and other facets of system medicine. For instance, NetDx utilizes Cancer Genome Atlas data to predict survival rates across various tumor types, each represented as a Patient Similarity Network (PSN) [172]. Table 9 summarizes patient similarity research focus and data sources.

Table 9: Patient Similarity Research Focus and Data Utilized

Focus	**Data Utilized**	**Reference**
GBM, Cancer	mRNA, DNA methylation, miRNA	[187]
Hepatocellular carcinoma	EHR after Transarterial Chemoembolization (TACE)	[191]
Diabetes, schizophrenia, cancers	EHR	[186]
Diabetes onset prediction	EHR	[188]
Stroke, ischemic heart disease, cerebrovascular disease	Chinese stroke data center	[209]
General concept	Genomics, proteomics, system medicine	[24]
Cancer survival	Cancer Genome Atlas data	[172]

2.2.5 *Performance Evaluation of the Existing PSNs*

In comparing various methods, DeepPatient [186] outperformed other feature learning techniques, achieving a high accuracy of 93%. Notably, it surpassed PCA, K-means, Gaussian mixture model (GMM), and independent component analysis (ICA) that achieved accuracies ranging from 87.9% to 92%. In another study, multidimensional patient similarity [178] using a

supervised approach achieved an accuracy of 77%, while hierarchical and K-means methods followed with accuracies of 73% and 71%, respectively.

2.3 Data Quality Management

Data quality is defined by Wang and Strong [210] as "data that are fit for use by data consumers". High-quality data should be fundamentally desirable, contextually relevant for the purpose, and easily available to data consumers. "The state of a set of values of qualitative or quantitative variables" is referred to as data quality by Prichard [211]. It is a judgment or evaluation of data's suitability to fulfill a certain function in a given environment, such as operations, decision-making, or planning. Measuring data quality has to be done with a goal in mind. Measurement is necessary for data management, and when used to unknown data, it's goal is discovery and definition. When applied to data with known problems, its goal is to create a foundation for improvement. Its challenge is to preserve quality when applied to reliable data. Measurement can also be applied to data that hasn't had any issues in order to control the risks of DQ that come with such data [212].

2.3.1 Dimensions of Data Quality

A Data Quality dimension [213] is a term used by data management experts to describe a data characteristic that may be evaluated or scored against established criteria to determine data quality. Identifying which data items need to be examined for data quality is a common Data Quality Assessment technique. We need to define values or ranges that indicate both excellent and unsatisfactory data quality for each data quality dimension. Each dimension is likely to have a different weightage, and in order to get an accurate assessment

of data quality, the system must figure out how much each dimension contributes to the overall data. The essence of evaluating data quality is to provide a framework that can be utilized for stating expectations, offering a mechanism for quantification, defining performance targets, and implementing the supervision process to verify that stakeholders adhere to the policies [214].

Table 10: Data Quality Dimensions

Dimension	Definition	Measure	Measure Type	Related Dimension
Completeness	The extent to which data is present compared to what is expected to be present.	A metric for the existence of non-blank values or the absence of blank (null or empty string) values.	Assessment	Validity, Accuracy
Uniqueness	Ensuring each item is recorded only once, regardless of how it is recognized.	Analysis of the number of items recorded in a data collection compared to the number of items present in the real world.	Discrete	Consistency
Timeliness	The extent to which data accurately reflects reality at the specified time.	Difference in time	Assessment and Continuous	Accuracy
Validity	Data is acceptable if it adheres to defined syntax (format, type, and range).	Comparison of the data to the metadata or documentation for the data item.	Assessment, Continuous, Discrete	Accuracy, Completeness, Consistency, Uniqueness
Accuracy	The degree to which data faithfully represents the "real world" item or event under consideration.	The degree to which the data reflects the qualities of the objects it represents in the actual world.	Assessment, Continuous, Discrete	Validity
Consistency	Ensuring no differences when comparing two or more representations to a definition.	Pattern and/or value frequency analysis.	Assessment and Discrete	Validity, Accuracy, Uniqueness

The six core dimensions [213] of data quality are: 1) Completeness, 2) Uniqueness, 3) Timeliness, 4) Validity, 5) Accuracy, and 6) Consistency which are detailed in Table.10.

It's crucial to remember that these data quality requirements aren't always satisfied 100 percent of the time; for example, data might be accurate but incomplete, or it can match all criteria except timeliness.

Dimensions are extremely context-dependent, and their relevance and value can differ significantly among organizations. As decision makers must make choices based on data, it is critical to conduct a quick data audit prior to assembling key performance indicator (KPI) findings in a performance report based on quality dimensions [211].

2.3.2 Data Quality Improvement Methods

Data quality improvement methods uses a variety of approaches to enhance data quality, including algorithms, heuristics, and knowledge-based activities. This section details few of the DQ improvement practices in literature.

Methodologies for data improvement include two types of strategies: data-driven and process-driven [215]. Data-driven methods increase data quality by changing the value of data directly. Process-driven strategies enhance quality by rethinking data creation and modification procedures. Data standardization [216] is frequently used in the master-data (data as it is gathered from various sources and organized into a single repository) domain, which has stringent data quality standards. Name standardization, address standardization, and product standardization are all common approaches. To find duplication, such as two customer records that pertain to the same person, matching algorithms are employed. Typically, a data standardization phase is performed prior to matching, and data consolidation techniques are used to

compare information in order to detect duplicates. Data enrichment [216] is the process of supplementing a record's known properties with new data obtained from reliable internal or external sources. The practice of combining two or more duplicate records into one record is known as data consolidation [216], or data merging. Extract-Transform-Load (ETL), often known as data integration [216], is a vast range of methods for moving data from one system to another. The data must first be extracted from the source system, and then transformations must be performed before the data can be loaded into the destination system. Luzzu [217] is a methodology for assessing the quality of linked data and includes an interface for establishing new quality measures as well as an ontology-driven back end for expressing quality metadata that may be reused across many semantic frameworks.

Another widely used tool is DaQL (Data Quality Library), which is used to continually assess the quality of data in order to improve the prediction accuracy of ML models. DaQL is first proposed in [218] and further enhanced to DAQL 2.0 [219]. The main enhancement was to create data quality standards for complex data objects or data entities, rather than considering a single data item and its relationships.

For enhancing data quality, the authors of [220] suggest a framework of data profiling based on DL and statistical model techniques. Data profiling evolved as a set of methods for statistical analysis and quality evaluation of data values inside a dataset. A data profiling tool generates a frequency distribution of the possible values for each column in a table. The article "The practitioner's guide to DQ improvement" [221] provides a thorough explanation of the methodologies utilized in the process of data profiling.

2.4 Resource Awareness

The optimization of energy consumption in continuous data-flow applications for IoT devices presents a significant challenge, as noted in various research works [145, 151, 222]. Anomalies occurring within the network infrastructure can substantially escalate energy consumption levels. To enhance the efficiency of continuous data flow applications, it is advisable to employ mathematical programming, such as a lightweight anomaly detection method, to evaluate node reliability [223]. Among the innovations aimed at optimizing edge resources, the Quality of Edge Computing Services stands as a pivotal metric, influencing the efficacy and responsiveness of applications in real-time environments [39].

Innovative hardware solutions have also emerged to tackle the energy efficiency challenge. For instance, a hardware prototype developed on an IoT-based microcontroller platform [145], effectively compresses real-time electrocardiogram (ECG) signals through sparse time-frequency domain encoding. Additionally, advancements in the form of an energy-conscious cyber-physical therapy system (T-CPS) framework [146] have been reported. This framework harnesses multimodal sensing capabilities to offer cost-effective and energy-efficient healthcare services, contributing to the ongoing efforts to optimize energy consumption in IoT applications.

Although some studies highlight resource-aware federated learning, very few works explored federated learning resource profiling that aids in node selection for optimizing the overall performance of the federated learning context. The following subsections review existing works that considered

resource characteristics for FL, FL inspired resource profiling, and FL driven client selection.

2.4.1 Resource-aware Federated Learning

Recently several research works related to federated learning have focused on developing algorithms for resource optimization based on one or more criteria including for instance cost, time, energy consumption, CPU power, and memory, using different algorithms. In [224], authors developed a deep reinforcement learning algorithm using an optimization function for minimizing the cost based on training time and energy consumption. The objective is to increase the speed of training while conserving the energy consumption of mobile devices by controlling the CPU cycle frequency. Furthermore, the authors of [225] proposed an optimization framework based on reinforcement Q-Learning. Their goal is to minimize the communication rounds and maximize accuracy. A hierarchical game framework was proposed to allow dynamic edge association and resource allocation in self-organizing Hierarchical FL networks [226]. The authors of [227] introduced the q-Fair optimization objective to emphasize fairness distribution among edges in FL according to performance which is based on the model's accuracy. To minimize FL training latency, the authors of [228] proposed a Multi-Armed Bandit algorithm based on training latency, availability, and fairness constraints. A framework proposed in [229] describes a resource and data-aware FL approach that characterizes the agents/resource according to certain criteria. Initially, it uses the local data size to classify the resources. Other criteria can also be used for resource selection to more accurately classify the resources used in a priority-based decision-making model such as

MCDM. Experiments show improvement in accuracy with lower training time over the standard FL baseline. The authors of [230] studied the effect of resource and dataset heterogeneity on the training time of FL deployed on AWS EC2. Their experimentation implied a severe impact of resource heterogeneity; in terms of response time and data quality; and on the training time of FL.

2.4.2 Federated Learning Inspired Resource Profiling

Few resource-aware FL frameworks adopted resource profiling to improve the performance of the FL model. The authors of [231] proposed a Tier-based FL system that classifies clients adaptively into groups according to their performance, then only select the nodes within the same group to avoid straggler problems; delays caused by the slowest devices, resulting from resource heterogeneity. The results from their experiments showed that the proposed approach improved the training speed while keeping the same or increasing the accuracy. Furthermore, the authors of [232] adopted a low-complexity profile-based resource-aware allocation strategy in FL called Dispersed Federated Learning (DFL). They implemented an optimization problem for resource allocation and device association by defining a preference and resource profiles. Preference profiles are used to allocate resource blocks to devices having a lower cost. Alternatively, the authors of [233] proposed a FL time minimization problem to control the data size used in each device using the makespan minimization problem and assign identical tasks to heterogeneous resources. They adopted an optimal polynomial-time scheduling algorithm called OLAR.

2.4.3 Federated learning Driven Edge Node Selection

In FL, the number of clients may be large, but the bandwidth available for model distribution and re-upload is limited, making it advisable to only involve only few of the edge nodes in the training process at any given time. This is a challenge that must be overcome to ensure that the model training process runs smoothly and without delay. With regards to training efficiency, quality of the final model, and fairness, the client selection policy is crucial to a FL process. In [234], the authors propose a model wherein the Lyapunov optimization problem of selecting clients with the highest degree of fairness is handled. Although FL has the potential to safeguard privacy in distributed ML, its limited applicability is due to inefficiencies in its communication [235]. With hierarchical FL (HFL), workers who own the data first send their updated model parameters to the edge servers for intermediate aggregation before sending everything to the central server for global aggregation. For effective HFL deployment, it is necessary to address edge association and resource allocation for non-cooperative parties such as edge servers and model owners. Using the edge server's predicted bandwidth allocation management approach in self-organizing HFL networks, a hierarchical game framework [226] is proposed to analyze the dynamics of edge association and resource allocation. While collecting more local data from clients enables more precise global model building, the resulting heavy data arrivals at the edge could be counterproductive to queue stability. The authors in [236] propose an algorithm that varies the number of clients chosen based on their current resource usage for transmission in order to maximize the time-averaged federated learning accuracy while maintaining queue stability.

In a traditional wireless network, an FL setup would involve clients sharing a wireless connection to a central server where the federated models would be trained. Due to the fact that both radio energy and client energy are limited [237], the learning performance is dependent on how clients are selected and how bandwidth is distributed among the selected clients in each learning round. To formulate a stochastic optimization problem for joint client selection and bandwidth allocation under long-term client energy constraints, the article [238] developed a new algorithm that uses only currently available wireless channel information but can still guarantee long-term performance.

2.5 Research Gaps

To the best of our knowledge, the following are the major conclusions drawn from the works included in the state-of-the-art literature review:

2.5.1 Smart Connected Health

- Many studies have examined the transition from siloed to SCH systems from the perspective of the patient experiences. In spite of this, there is a lack of a literature-based classification of SCH.
- In addition, the current state of knowledge on SCH lacks an architecture for SCH systems, complete with SCH building blocks that capture technological aspects, the environment, and the involved stakeholders.

2.5.2 Patient Similarity Network

- The state-of-the-art literature provides a wealth of options for constructing PSN, from more conventional ML methods to cutting-edge DL approaches.

- Clinical narrative data exhibits diversity and heterogeneity, but it contains hidden information that provides valuable insights for identifying patients with the highest similarity to each other. Knowledge of relevant historical information is crucial for understanding complex medical contexts, given the time-sensitive nature of medical events. Existing PSN approaches often aggregate medical data into a time-based "vector" representation of the patient, losing crucial temporal context. Research shows that the interpretation of temporal representation in noisy clinical datasets is challenging, leading to inaccurate outcome predictions.
- Health datasets such as EHRs are high-dimensional and varied.. There's limited research on using hybrid generalized models combining ML, NLP, and autoencoders to manage data heterogeneity and address big data challenges in PSN.

2.5.3 Data Quality

- In healthcare, low DQ at the edges presents a significant challenge that can hinder reliable and precise decision-making. A small number of studies looked at how poor edge data quality affected model accuracy, but even fewer looked at how incorporating FL ideas into data quality could improve accuracy.
- Never before has the literature investigated data quality-aware edge selection and profiling for PSN, or its integration with FL services to address faulty data issues and privacy at dispersed client sources.

2.5.4 Resource Awareness

- Resource-aware FL has been the subject of some research, but resource profiling to aid in node selection and improve federated learning's overall performance has received comparatively little attention.
- There are no existing literature on federated learning resource profiling that combines both edge static and dynamic resource quality profile attributes with edge reputation information to ensure optimal edge node selection.

In light of these knowledge gaps, we set out to develop optimal approaches for handling the heterogeneity of data in PSN, taking temporal medical data into account, ensuring high-quality data is aggregated at the network's edges, evaluating resources at the edge using multiple criteria, and establishing federated efforts to decrease response times and improve prediction performance.

Chapter 3: Research Methodology

To begin, we utilized the scoping research review technique to classify and analyse existing literature that reports on the implementation, design, and solutions of SCH in healthcare in order to address few of the research questions in SCH. Further, we propose our PSN Multidimensional Data Fusion model and PSN Data Quality Management Model based on data quality profiling. Finally we propose the FL-inspired resource management model.

In order to respond to the crisis in the new post-pandemic reality, the authorities in the healthcare domain must review the current healthcare systems [239], and implement new care models to address primary, secondary, and acute care. In addition to demanding more flexible, interprofessional care delivery with empowered frontline staff utilizing technology, patients are also demanding continuous access to healthcare services in a safe and convenient manner. This calls for a new framework in order to help healthcare organizations coordinate the numerous interrelated changes needed to support virtual care.

Redrafted from: A. N. Navaz, M. A. Serhani, H. T. El Kassabi, N. Al-Qirim, and H. Ismail, "Trends, Technologies, and Key Challenges in Smart and Connected Healthcare," in IEEE Access, vol. 9, pp. 74044-74067, 2021, doi: 10.1109/ACCESS.2021.3079217.

3.1 SCH : Trends, Architecture, and Case Study

Examining the current state of smart connected health technologies and their effects on healthcare systems is the goal of this scoping review. The objective of this study is to address the following research: "How can we create intelligent SCH systems and strategies that deliver efficient, proactive, patient-centric healthcare?" To provide insights into this question, we propose the following contributions.

- Present a classification model for existing SCH models, solutions, and architectures.
- Introduce an architectural model that encompasses the technical aspects, environment, and key stakeholders of SCH solutions. This model incorporates innovations like IoT, Big Data analytics, and AI to enhance healthcare system efficiency.
- Conduct a case study on how SCH was employed during the COVID-19 pandemic, showcasing how various SCH technologies were used to identify, monitor, and control the virus. The study also emphasizes the optimal allocation of technical resources to combat the pandemic.
- Identify major SCH challenges and propose future research and technology directions for next-generation SCH. Additionally, provide recommendations for future SCH system implementations based on lessons learned from the pandemic.

3.1.1 SCH Architecture

We present an SCH architecture and examine a cardiovascular disease (CVD) monitoring implementation scenario. We outline the technological components and potential applications stemming from this solution.

Data is at the core of SCH, with diverse technologies used for data collection, processing, interpretation, and visualization to facilitate data-driven decision-making. The primary challenge is not data provision but rather its interpretation to derive valuable insights. Hospital records, containing extensive historical patient data and medical records, represent a valuable resource for analysis. Analyzing this data can reveal correlations, patterns, and trends that aid in predicting disease occurrences and informing treatment decisions. Deep-learning algorithms outperform humans in trend detection when analyzing a patient's historical data [44].

3.1.2 SCH Building Blocks

In this section, we present a comprehensive outline of SCH building blocks, encompassing the technological aspects of pandemic solutions, their applicable environments, and the primary stakeholders involved. The utilization of these SCH building blocks is context-specific, depending upon the application's purpose and the primary beneficiaries. As depicted in Figure 5, after setting up the environment, the process aligns with the big data value chain approach, including data ingestion, preprocessing, processing and storage, analytics, and ultimately, visualization and insight delivery. The proposed SCH architecture stands out due to the incorporation of smart solutions at various stages.

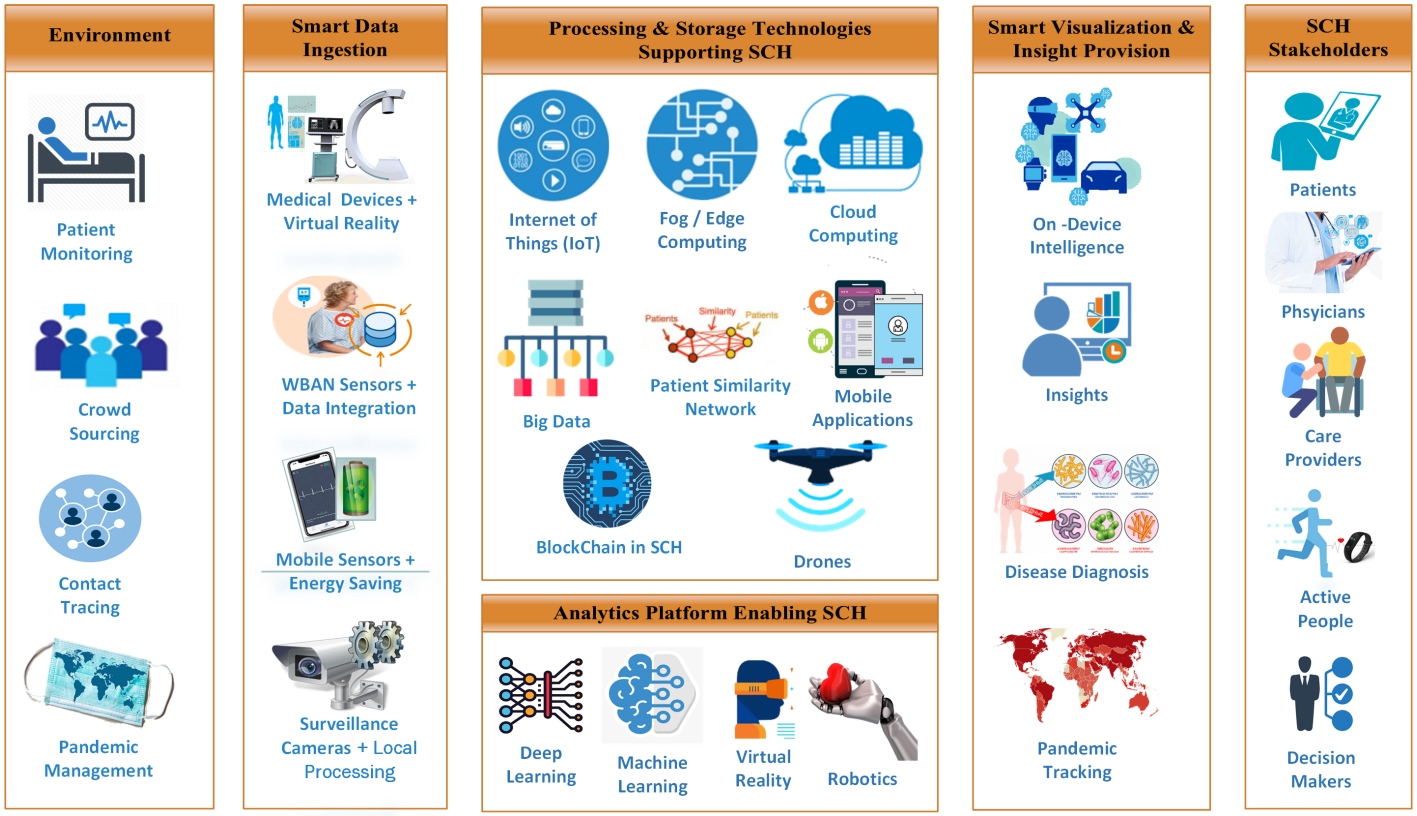


Figure 5: Building Blocks of Smart and Connected Health (SCH) Architecture

3.1.3 SCH Building Blocks Application Scenario

The world's leading cause of mortality is CVD, and electrocardiography (ECG) is a widely used method for measuring cardiac activity and detecting heart defects [240]. An illustrative scenario involves a patient undergoing heart disease monitoring in a healthcare facility. Here, ECG data is collected via wired or wireless biosensors. In ambulatory or home settings, the data is transmitted to a mobile device equipped with efficient data ingestion capabilities. Any ECG signal irregularities, such as fluctuations, are identified during processing, conducted within Fog, Edge, and Cloud Computing environments. During this stage, the ECG signal channels undergo translation into numeric sequences, preceded by preprocessing steps (e.g., cleaning, noise removal, filtering), feature extraction, and selection. DL techniques can predict signal anomalies, aiding in the early detection of conditions like myocardial infarction, arrhythmia, and heart failure, facilitated by analytics-integrated Smart Connected Healthcare (SCH) platforms. These platforms also enable real-time interactive visualizations, insights delivery, and recommendations accessible via mobile devices and web applications for stakeholders. To ensure data security and privacy, patient ECG data is securely stored in big data platforms with blockchain-based trusted transaction features. Additionally, intelligent features encompass sensor instrumentation for edge data preprocessing, device energy harvesting, self-adaptation, and self-learning algorithms, all contributing to swift emergency responses and patient risk mitigation [240].

3.1.4 Case Study on SCH: COVID-19 Pandemic Management

We provide an in-depth case study on China's utilization of SCH technologies during the COVID-19 pandemic, showcasing the effectiveness of each technology in managing and containing the outbreak.

In Figure 6, we depict the timeline of SCH adoption for managing the COVID-19 pandemic in China. This curve illustrates the rapid and significant impact of technologies like AI, big data, robotics, IoT, drones, and mobile applications in swiftly flattening the curve, resulting in a decline in new cases within just two months.

In summary, this research lays a strong foundation for the study of SCH, paving the way for future advancements and innovative solutions in the field.

Redrafted from: A. N. Navaz, H. T. El-Kassabi, M. A. Serhani, A. Oulhaj, and K. Khalil, "A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine," Journal of Personalized Medicine, vol. 12, no. 5, p. 768, May 2022, doi: 10.3390/jpm12050768.

This research work is based in [full or part] on the previously published article(s) listed above. I have permission from my co-authors/publishers to use the work listed above in my thesis/dissertation.

3.2 PSN Multidimensional Data Fusion Model

PSNs are designed to manage heterogeneous data by converting each datatype to a similarity network and then easily integrating/aggregating them into one similarity network using, for example, a fusion algorithm.

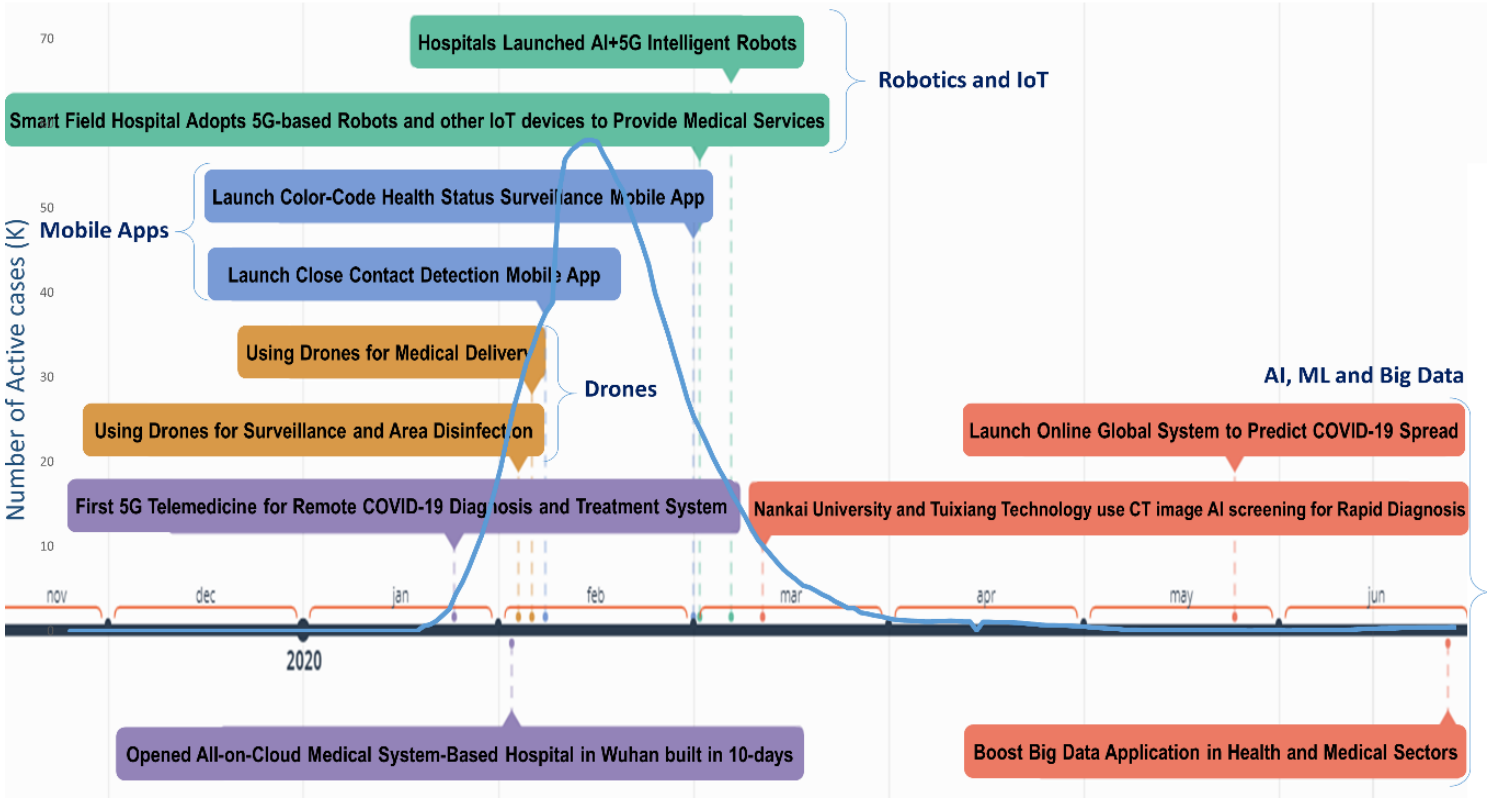


Figure 6: Timeline of China's SCH Adoption to Flatten the Curve of COVID-19

Patient similarity analysis is a tool used in precision medicine to better predict the health outcomes of individual patients. The patient similarity network (PSN) model allows for accurate and generalizable classifiers by giving classifiers the ability to naturally incorporate heterogeneous data and handle missing information [26].

3.2.1 Multidimensional Data Fusion Model Architecture

In this section, we describe the proposed system architecture in which a DL-based approach is adopted for building patient similarity. We emphasize the main processes involved in implementing our solution, including the data-collection phase, DL model development, training, testing, model accuracy evaluation, and diagnostic prediction and clinical recommendations.

3.2.2 PSN Data Fusion Model Formulation

We propose a model formulation to represent patients and derive a similarity measure based on the vectors generated from medical events. To model this data, we denote the patient set as $S= s_1,s_2,\ldots,s_n$, where s_i is the vector of the i^{th} patient, and n is the number of patients. This vector comprises a tuple of two main parts, namely, the static part St and dynamic part D, $s_i= (st_i,d_i)$.

3.2.2.1 Static Data Modeling

The static data part St represents the patient's profile information containing age, gender, multiple laboratory test items, and multiple disease diagnoses. Further, we model the similarity between two of the selected features, age and diabetes, to illustrate this point.

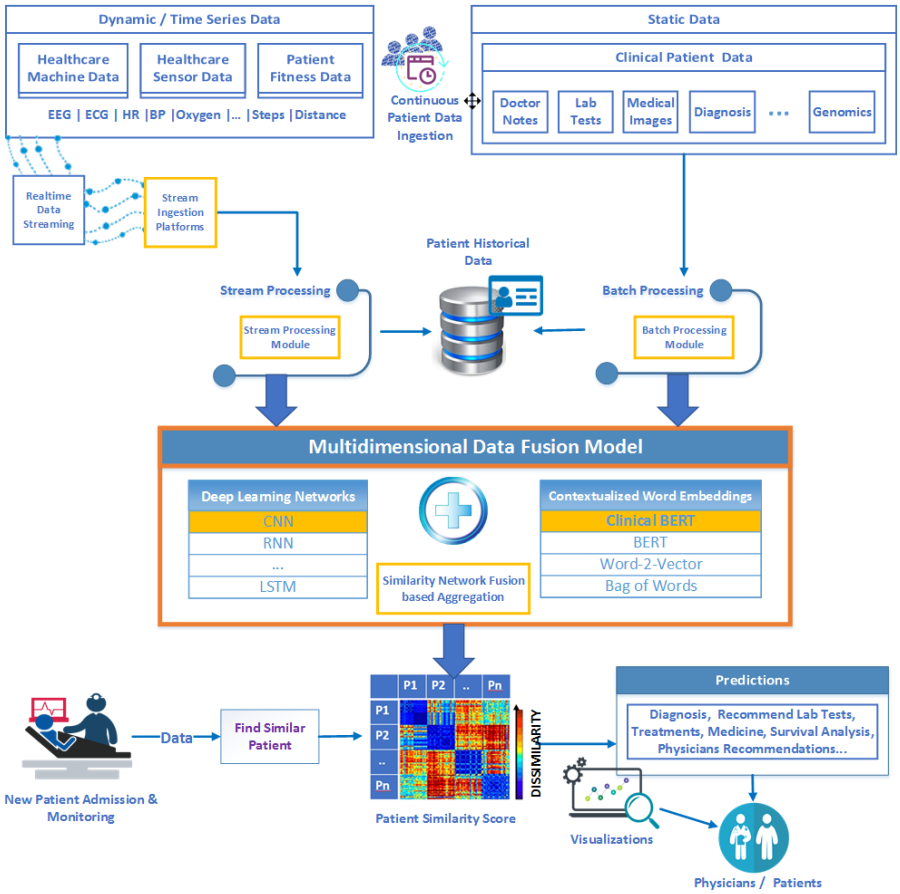


Figure 7: Multidimensional PSN Data Fusion Model

A. Feature Similarity for Age

We denote age_i and age_j as the ages of patients i and j, respectively. We can represent the feature similarity fs^1 for age as the ratio of the smaller age to the larger age [241].

$$fs^1_{(i,j)} = min(age_i, age_j)/max(age_i, age_j) \tag{3.1}$$

B. Feature Similarity for Diabetes

Other static features include events, such as patients having a chronic disease, represented as a Boolean value. For example, when a patient is diabetic, we define the similarity feature fs^2 between patients i and j as 1 if both patients have the same condition (either both diabetic or both nondiabetic) and 0 otherwise.

$$fs^2{}_{(\mathrm{i,j})} = \begin{cases} 1, & \text{if } \mathrm{diab_i} = \mathrm{diab_j} \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

C. Global Static Patient Similarity

We calculate the global patient similarity for static features using the following weighted sum of all the static feature similarities as a single measure of static patient similarity (STPS) for patients i and j. We use a weight vector $\mathrm{WV} = \mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_{\mathrm{nf}}$, where nf is the number of static features used to evaluate the patient similarity, $\mathrm{w_k}$ is the weight given for each static similarity feature $\mathrm{fs^k}$, $\mathrm{w_k} \in W, and \sum_{K=1}^{nf} w_{\mathrm{k}} = 1$.

$$STPS_{(\mathrm{i,j})} = \sum_{K=1}^{nf} w_{\mathrm{k}} fs^k{}_{\mathrm{i,j}}, \text{where } k = 1, 2, 3, \ldots, nf. \tag{3.3}$$

3.2.2.2 Dynamic Data Modeling

The dynamic data part D is extracted from the EHR data, which is a time-series vector representing the number of visits m and is denoted by a sequence of visits as $\mathrm{D} = \mathrm{PVd_1}, \mathrm{PVd_2}, \ldots, \mathrm{PVd_m}$. Each visit $\mathrm{PVd_i}$ is denoted

by a high dimensional vector $PVd_{\mathrm{i}} \in R$, where each element indicates that the patient has a medical event value represented as a real number. Therefore, the horizontal axis indicates the rows (i), each of which represents a visit PVd_{i}, and the vertical axis indicates the columns (j), which represent the medical events $x_{\mathrm{i}} \in X$, where X is the set of medical events, that is, features. The $(\mathrm{i,j})^{\mathrm{th}}$ value is observed at time t_{i} of PVd_{i} for a certain patient. The number of visits varies for different patients. Thus, the dimension of this matrix is defined as dim = $max(D)_{i=1}^{m}$. This variable-sized data can be managed using an autoencoder-based long–short-term memory (LSTM), which is detailed in the following section.

A. LSTM

LSTM is a variation of deep RNNs that have been commonly adopted in diverse domains such as language modeling and speech recognition [242]. It overcomes vanishing gradient problems using a forget gate that allows the error to be backpropagated [213]. In our model, LSTM is a gated RNN with an input vector, which is the dynamic part vector $d_i \in R^S$ of the patient's set PV.

B. Patient Visit Matrix Embedding (Data Dimension Reduction)

The dynamic data part D is fed into one layer of the time-series LSTM model encoder to preserve the temporal features of patients' data. This layer reduces the data dimension to produce an output vector D', which includes embeddings of a smaller dimension d as the final hidden state. This is performed to reduce data dimensions and learn relationships among features. Consequently, each visit PVd_{i} is mapped into an embedding matrix $EB_i \in R^{\mathrm{dim}}$, where $dim < |X|$ is the embedding dimension, with X, representing the orginal

number of features. Using the rectified linear activation function (ReLU) in (3.4), the summed weighted input is transformed into an output using a formula similar to that in a previous study [182], where $W_v \in R^{\dim * |X|}$ and $b_v \in R^{\dim}$ are the weight matrix and bias vector to be learned, respectively.

$$e_i = ReLU(W_v PV d_i + b_v) \tag{3.4}$$

$$ReLU(x) = max(0, x) \tag{3.5}$$

This operation results in an embedding matrix EB_i for each patient, resulting in a lower feature dimension than the original dataset.

C. Similarity Network Fusion

SNF is a new nonlinear computational approach for integrating and fusing different PSNs [187]. First, we normalize each matrix by dividing each row element of the matrix by the sum of the rows so that the sum of all the elements in each row is 1.

$$w_{i,j} = \frac{m_{\mathrm{i,j}}}{\sum_{\mathrm{j=1}}^{\mathrm{n}} m_{i,j}}, \tag{3.6}$$

where $w_{i,j}$ is the normalized value of each element $m_{i,j}$ of the similarity matrix. Then, the normalized matrix W can be symmetrized as

$$W_{Sym} = (W + W^T)/2, \tag{3.7}$$

where W^T denotes the transpose of W. The resulting matrices are defined as STM and DM to represent the static data similarity matrix and dynamic data similarity matrix, respectively. Next, we use the K-nearest neighbor method to calculate the local similarity for each matrix [187].

$$w_{\mathrm{i,j}}^{,} = \begin{cases} \frac{w_{i,j}}{\sum_{y \in W} w_{\mathrm{i,j}}^{\mathrm{y}}}, j \in N \\ 0, otherwise \end{cases} \tag{3.8}$$

In Equation 3.8, we define the normalized weight, where N represents the set of K nearest neighbors of patient i obtained from both matrices, and within this set, y an individual nearest neighbour. In this context, the hyperparameter K, is utilized in local similarity computations, and is determined by the user. Thus, the strongest links with the highest weights are selected, and the weak links in the network are eliminated to reduce noise interference. Finally, the two updated matrices STM'^1 and DM'^2 will be fed to the SNF algorithm that will iterate for a given number of iterations T, starting at $MP^1_{\mathrm{t=0}} = STM$ and $MP^2_{\mathrm{t=0}} = DM$. In general, SNF fuses the similarity networks attained from different data types separately by aggregating their data. However, here, we use SNF to combine patient similarity matrices rather than raw data. Therefore, we modified the algorithm to aggregate the similarity values between each pair of patients into a single value in accordance with the following aggregation function based on weighted average [243].

$$MP^1_{\mathrm{t+1}} = \frac{wt_s STM^{'} + (1 - wt_s) MP^2_t}{2}, \tag{3.9}$$

$$MP^{2}_{t+1} = \frac{wt_d DM^{'} + (1 - wt_d) MP^{1}_{t}}{2}, \tag{3.10}$$

where wt_s and wt_d denote the weights according to the significance of each matrix estimated by the user. Here, MP^{1}_{t+1} is the state matrix transformed based on the STM similarity matrix after t iterations and MP^{2}_{t+1} is the state matrix transformed based on the DM similarity matrix after t iterations. In each iteration, the information of each similarity network is changed to produce two final state matrices that will be integrated into the fusion similarity matrix FM as,

$$FM = \frac{MP^{1}_{t} + MP^{2}_{t}}{2}, where\, t = T \tag{3.11}$$

This modification will distinctly indicate the strength of similarity between each pair of patients and reduce the noise and interference that can be attributed to the similarity of other patients. The integrated matrix, which was obtained from the aforementioned sequential operations, produces a PSN defined as G = (V, E). The vertex V represents the patient set S, and the edges E are weighted by the similarity level between the patients. The edge weights are denoted as a N × N similarity matrix FM resulted from the final iteration of the SNF algorithm, as explained earlier, where each element $w(i, j)$ indicates the similarity level between patients s_i and s_j.

Figure 8 shows the key processes associated with the building of a hybrid PSN, including static, dynamic, and fused similarity matrix constructions, as per the aforementioned formal description.

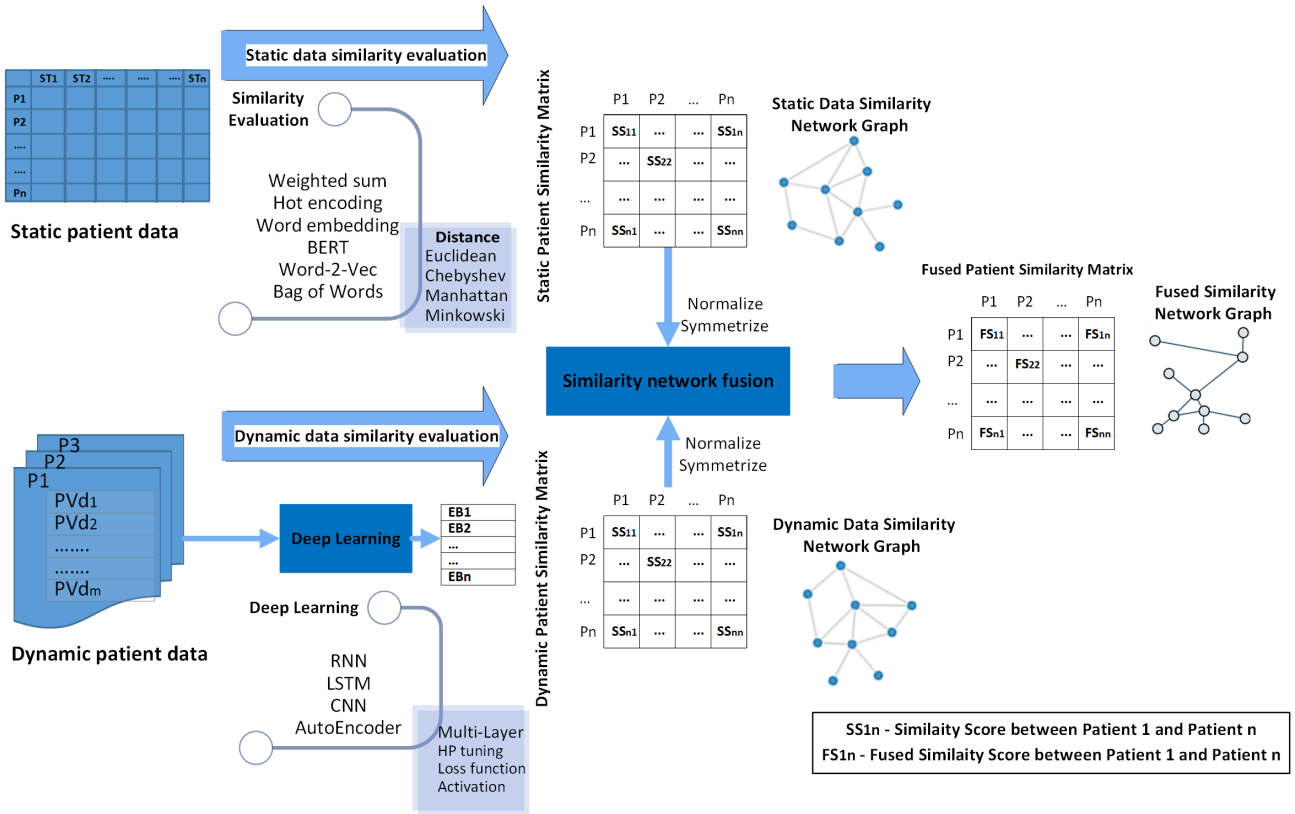


Figure 8: Key Processes in Building a PSN

3.2.3 PSN Construction Algorithms

In this section, we introduced three algorithms for constructing the proposed hybrid PSN. The first algorithm implements the procedure to generate the static similarity matrix, the second algorithm implements dynamic similarity matrix generation, and the third algorithm implements similarity matrix fusion.

Algorithm 1 outputs the STPS matrix based on the model explained in this study. The input to this algorithm is the static data part of the patient dataset, the list of selected features to be evaluated for similarity, the list of similarity utility for each selected feature, and the weights for each similarity feature.

Algorithm 1 Static data similarity evaluation algorithm

Input:
PList ▷ List of Patients
SFList ▷ List of selected features
SUList ▷ List of similarity utility for each feature
weights ▷ List of weight for each feature
Output:
SSM ▷ Static similarity matrix for all patients

1: **procedure** STATICSIMILARITYMATRIX($PList, SFList, SUList, weights$)
2: $SSM \leftarrow$ INITIALIZETOEMPTY
3: **for** $s_i \leftarrow 1, N$ **do** ▷ each patient i
4: **for** $s_j \leftarrow s_i + 1, N$ **do** ▷ each patient j
5: **for** $f_k \leftarrow 1, K$ **do** ▷ each selected feature (col)
6: $FSscore[s_i, s_j] \leftarrow$ GETSIMSCORE$(s_i, s_j, SUList[f_k])$
7: $SSM[s_i, s_j] \leftarrow SSM[s_i, s_j] + FSscore[s_i, s_j] \times weights[f_k]$
8: **return** *SSM*

Algorithm 2 applies the DL autoencoder model to generate the dynamic similarity matrix. It takes as input the dynamic data part of the patient list denoted as PV, the activation function, e.g., ReLu, the number of dynamic features to be evaluated, and the output embedding dimension.

Algorithm 2 Dynamic data similarity evaluation algorithm

Input:
DPList ▷ List of Patients with dynamic data
ACTF ▷ Activation function
NF ▷ Number of features
NEMB ▷ Embedding dimension
Output:
DSM ▷ Dynamic similarity matrix for all patients

1: **procedure** DYNAMICSIMILARITYMATRIX(*DPList*, *ACTF*, *NF*, *NEMB*)
2: $preprocess(DPList)$
3: $EB \leftarrow deepLearningAutoencoder(DPList, ACTF, NF, NEMB)$
4: **for** $s_i \leftarrow 1$ **to** N **do** ▷ each patient i
5: **for** $s_j \leftarrow s_i + 1$ **to** N **do** ▷ each patient j
6: $DSM[s_i, s_j] \leftarrow getSimilarityScore(EB[s_i], EB[s_j])$ ▷ Euclidean distance
7: **return** *DSM*

Algorithm 3 finalizes the fusion process and creates the fused patient matrix. It takes as input the two aforementioned matrices, the number of nearest neighbors K, and the number of iterations T required for the SNF process.

Algorithm 3 Similarity network fusion algorithm

Input:

STM ▷ Static similarity matrix

DM ▷ Dynamic similarity matrix

T ▷ Number of iterations to complete fusion

K ▷ Number of nearest neighbors

wt_s ▷ Weight for Static similarity matrix

wt_d ▷ Weight for Dynamic similarity matrix

Output:

$FPSM$ ▷ Fused patient similarity matrix

procedure SIMILARITYNETWORKFUSION(STM, DM, T, K, wt_s, wt_d)
  $M^1_{prev} \leftarrow STM$
  $M^2_{prev} \leftarrow DM$
  $normalize(STM, DM)$
  $symmetrize(STM, DM)$
  for $s_i \in STM$ **do** ▷ calculate local similarity for STM
    $neighborList \leftarrow nearestKNeihbors(s_i, K, STM, DM)$
    for $s_j \in neighborList$ **do**
      $STM[s_i, s_j] \leftarrow STM[s_i, s_j] / \sum_{i=1}^{K} neighborList[i]$
  for $s_i \in DM$ **do** ▷ calculate local similarity for DM
    $neighborList \leftarrow nearestKNeihbors(s_i, K, STM, DM)$
    for $s_j \in neighborList$ **do**
      $DM[s_i, s_j] \leftarrow DM[s_i, s_j] / \sum_{i=1}^{K} neighborList[i]$
  for $t_i \leftarrow 1$ **to** T **do**
    $M^1 \leftarrow (wt_s \times STM + (1 - wt_s) \times M^2_{prev})/2$
    $M^2 \leftarrow (wt_d \times DM + (1 - wt_d) \times M^1_{prev})/2$
    $M^1_{prev} \leftarrow M^1$
    $M^2_{prev} \leftarrow M^2$
  $FPSM \leftarrow (M^1 + M^2)/2$
  return $FPSM$

Redrafted from: A. N. Navaz, M. A. Serhani, H. T. El Kassabi, and I. Taleb, "Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling," Sensors (Basel)., vol. 23, no. 14, pp. 1–32, 2023, doi: 10.3390/s23146443.

This research work is based in [full or part] on the previously published article(s) listed above. I have permission from my co-authors/publishers to use the work listed above in my thesis/dissertation.

3.3 PSN Data Quality Management Model

When compared to more conventional classification algorithms, the proposed PSN fusion model improves classification accuracy when predicting a variety of patient health outcomes. The precision and efficiency of the PSN classification model can be boosted by assessing and enhancing the quality of the data. This can be accomplished by data profiling, which is the process of assessing data available from an existing information source (e.g., a database or a file) and compiling statistics or instructive summaries about that data. In addition, we present FDQP as an unique technique for ensuring data quality at the edge source.

3.3.1 Data Quality Profiling

Data profiling [244] is a collection of analytical approaches that assess actual data content to offer a full picture of each data element in a data source. The methodology used is data quality assessment and profiling all through the various stages of the data value chain as detailed in Figure 9.

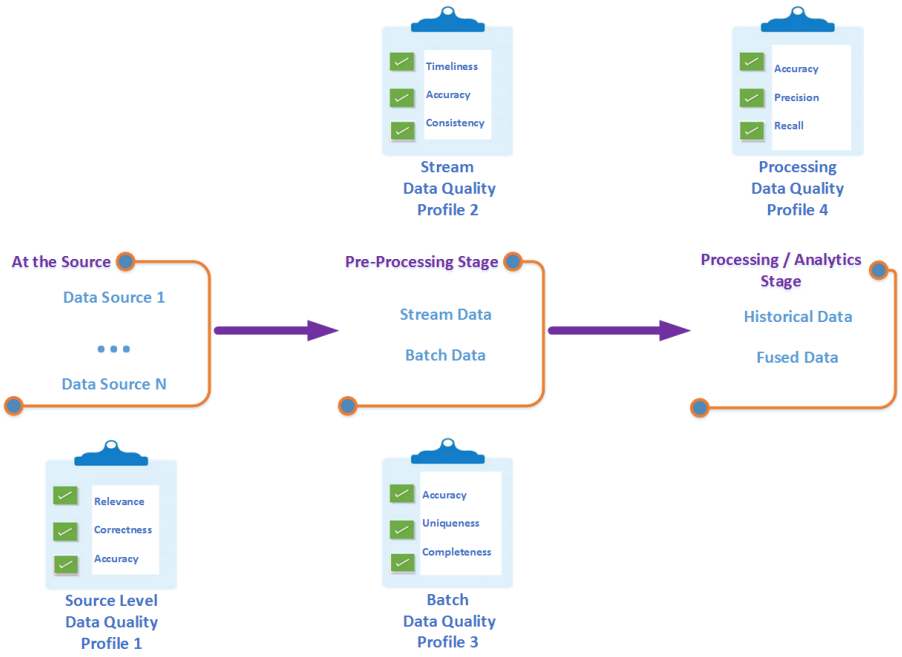


Figure 9: Data Quality Profiling at Various Stages of the Data Value Chain

To evaluate essential data quality characteristics, a set of data quality dimensions (DQD) are defined for each stage. Some data quality measures, such as Accuracy, are critical at each stage, while others may be stage-specific (source, pre-processing, processing/analytics).The DQP plays a pivotal role in our methodology. After some iterations, a basic DQP is transformed into a comprehensive DQP, with quality rules that ensure data integrity across all of its characteristics, dimensions, and stages.

Figure 10 illustrates the incorporation of the improved Data Quality Model into our extended PSN Multi-model. Throughout the Multidimensional fusion model's data acquisition, stream data pre-processing, batch data pre-processing, and processing phases, data quality is continuously evaluated and assessed.

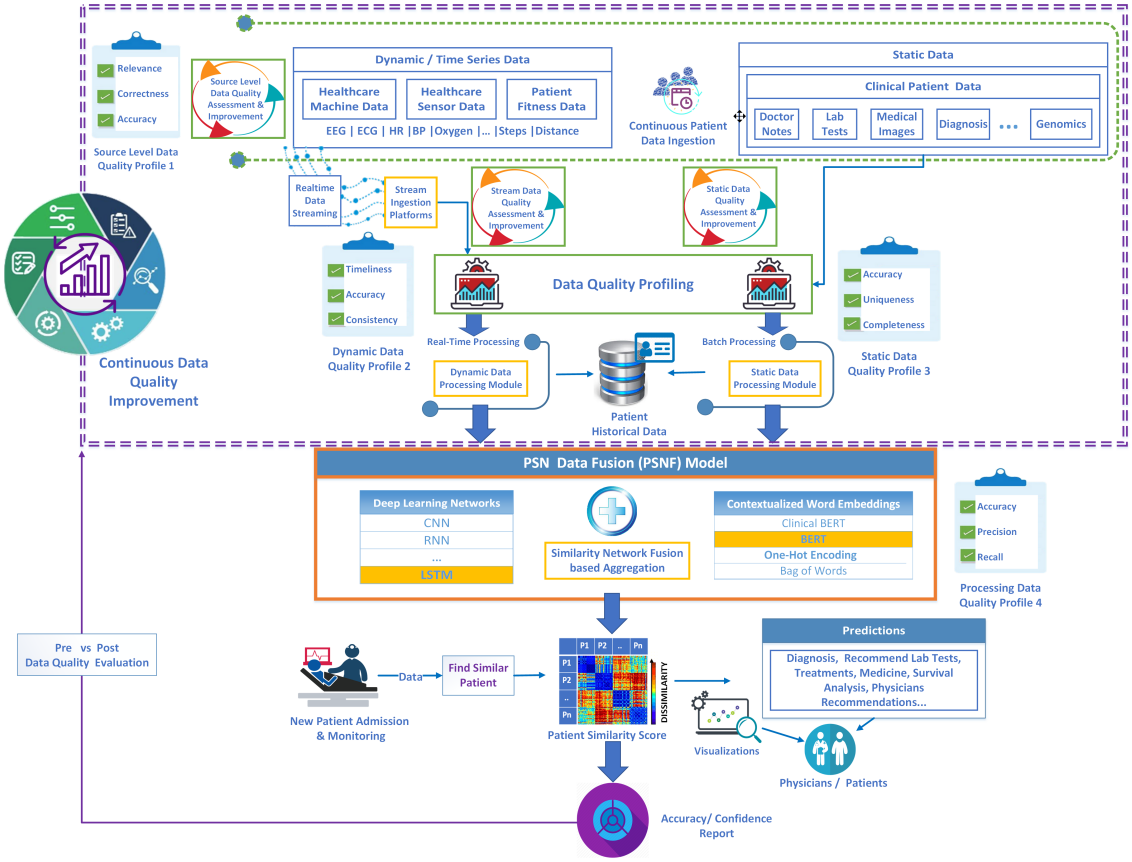


Figure 10: Enhanced PSN Multi-Model

In addition to a pre and post data quality evaluation based on prediction accuracy, a DQP is produced at each stage that is tailored to the data and processing workflows. Thus, the proposed end-to-end continuous data quality upgrade would unquestionably raise the PSN model's precision.

3.3.2 Data Quality Aware FPSN Model

The FDQP Model is a proposed approach to address privacy and data governance challenges at non-co-located data. It involves a centralized server pushing the basic DQP to the edge nodes, who then assess and evaluate the data quality in the relevant health datasets and update the DQP and send it back to the server as illustrated in Figure 11. Finally, the server combines the DQP's from the edges to build the FDQP, which serves as a basis for the Federated PSN.

An overview of the proposed model with FPSN and FDQP edge enhancement is shown in Figure 12. The objective of this framework is to enhance the precision and speed of data processing at the edge of the network. The architecture also features a cloud-based server that facilitates the profiling federation and the federated PSN score aggregation. Pre and post-data quality evaluation is performed, and the process is repeated until the data quality reaches acceptable tolerance levels.

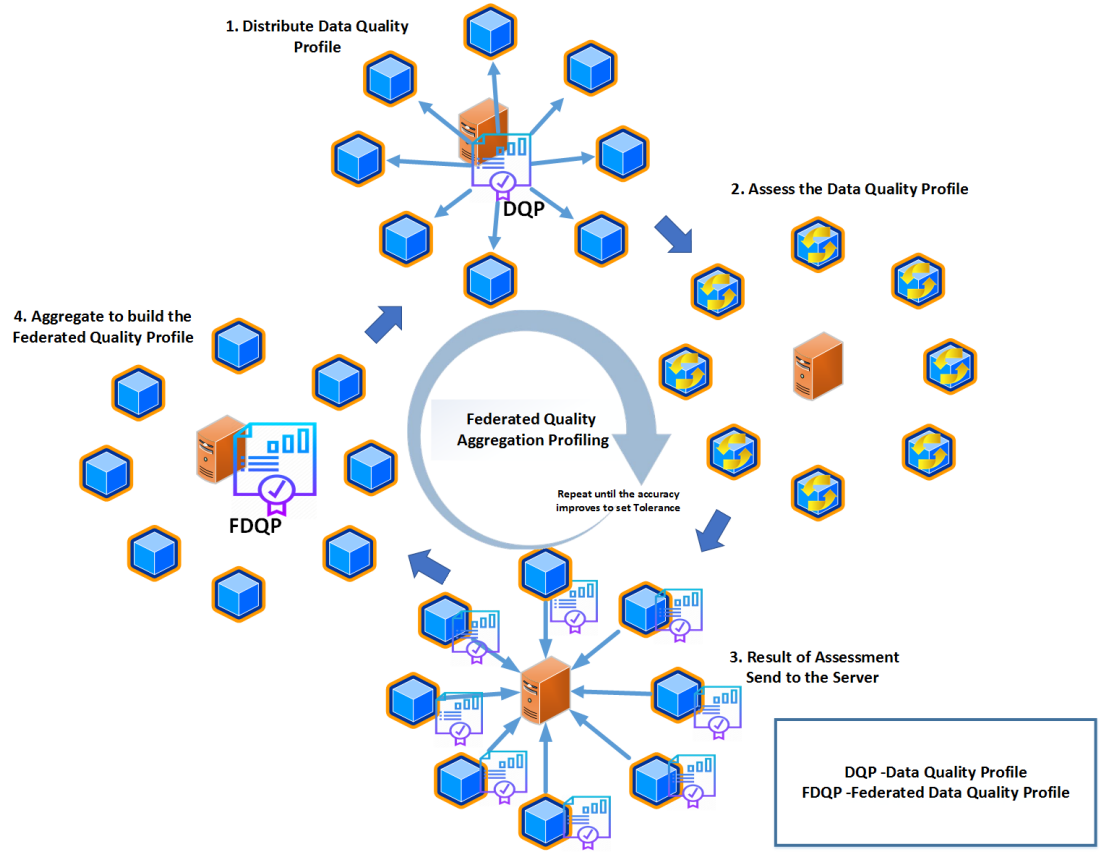


Figure 11: Federated Data Quality Profiling

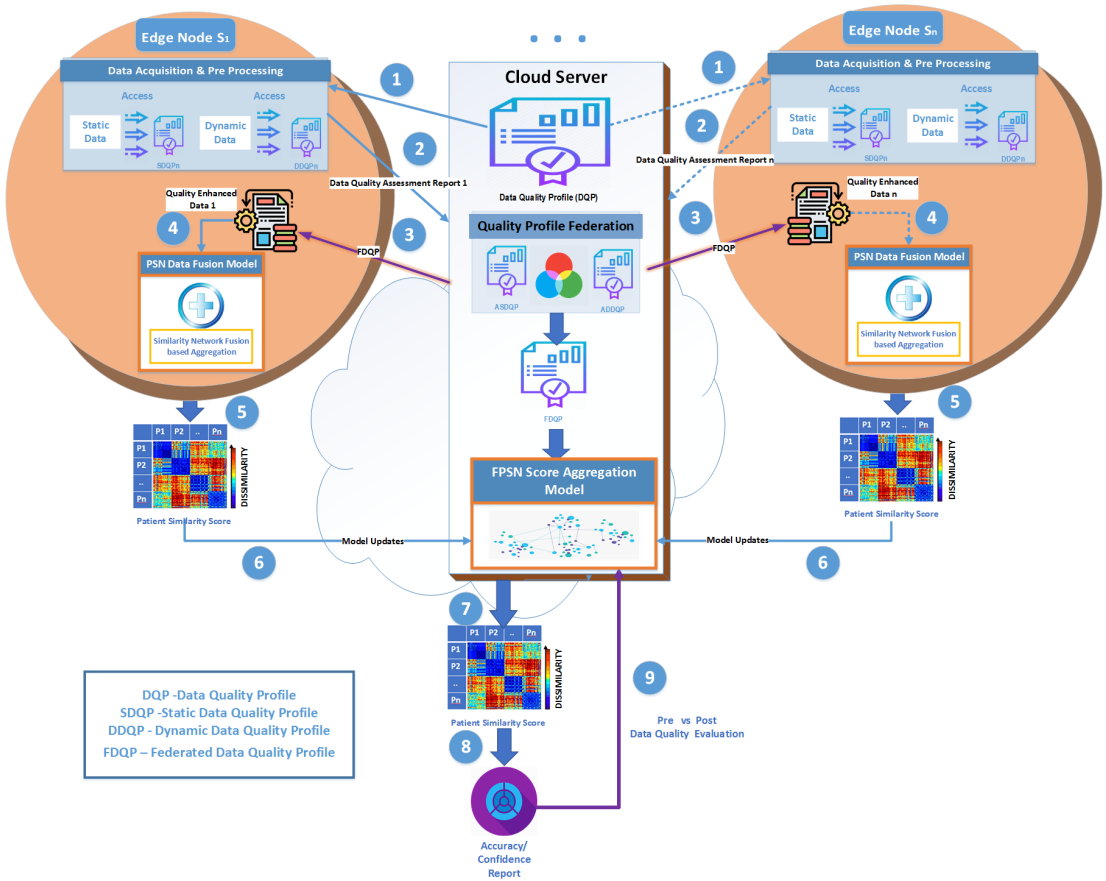


Figure 12: FPSN Enhanced by FDQP at the Edge

The following are the sequential steps of the model processes.

1. In the initial stage, the centralized cloud server sends the baseline DQP to the edge nodes.
2. Subsequently, at the edge local node, each node verifies and evaluates the data quality acquired from the sensors, updates the DQP, and transmits it back to the server.
3. Finally, the server integrates the DQPs received from the edge nodes to create the Federated Data Quality Profile (FDQProfile), which is transferred to the Edge nodes.
4. FDQProfile is then applied to the local edge data that creates the Quality enriched data, which will be the basis for The PSN Data Fusion model at the edge.
5. The resulting Patient Similarity Score is sent back to the cloud Server.
6. The FPSN Score Aggregation model receives the model updates from the edges and the aggregation of the similarity scores takes place at the cloud server.
7. The final Patient Similarity Score is used as a basis to detect the most similar patient.
8. DQ evaluation measures of performance such as accuracy is calculated.
9. The process is further assessed with pre and post-DQ evaluations.

3.3.3 Federated Data Quality Profiling Illustration

Figure 13 depicts Federated Data Quality (FDQ) profiling using an example. The baseline quality profile specifies the needed data quality characteristics such as completeness, accuracy, and timeliness, as well as the data quality standards. Based on the baseline quality profile, we identify the

dimensions of DQ with severe issues. A missing data-related rule, for example, will infer some actions (e.g., replace missing values with the mean) depending on the kind of data and the degree of tolerance specified in the DQP. This baseline DQP will be forwarded to the source edges, where the local dataset will be reviewed using the profile and a new quality profile will be constructed using some or all of the criteria. If the new profile satisfies the baseline quality profile, it is sent back to the server. It's worth noting that we're considering the edges of collecting and holding identical datasets with similar characteristics. Here, Figure 13 shows that source edge 1 with attribute ID 1 has a missing value of 70%, hence rule 2.2 is used to eliminate the whole column.

Similarly, at each edge node, a local profile is relayed to the server node, which aggregates all profiles. According to the context and the rules, the aggregate process will use min, max, total, or average aggregation. The aggregated quality profiles will then be integrated to generate the federated data quality profile, which will include optimal rules based on attribute correlations. Furthermore, the feature selection rules will be defined in the federated data quality profile based on attribute priority and ranking. Thus, the federated data quality profile will have a well-defined purpose, quantifiable metrics, and optimized rules for selection as well as explicit formulas for combining local/federated data profiles. Finally, the rules in the federated data quality profile are propagated and enforced across all nodes in the federation, ensuring enhanced data quality.

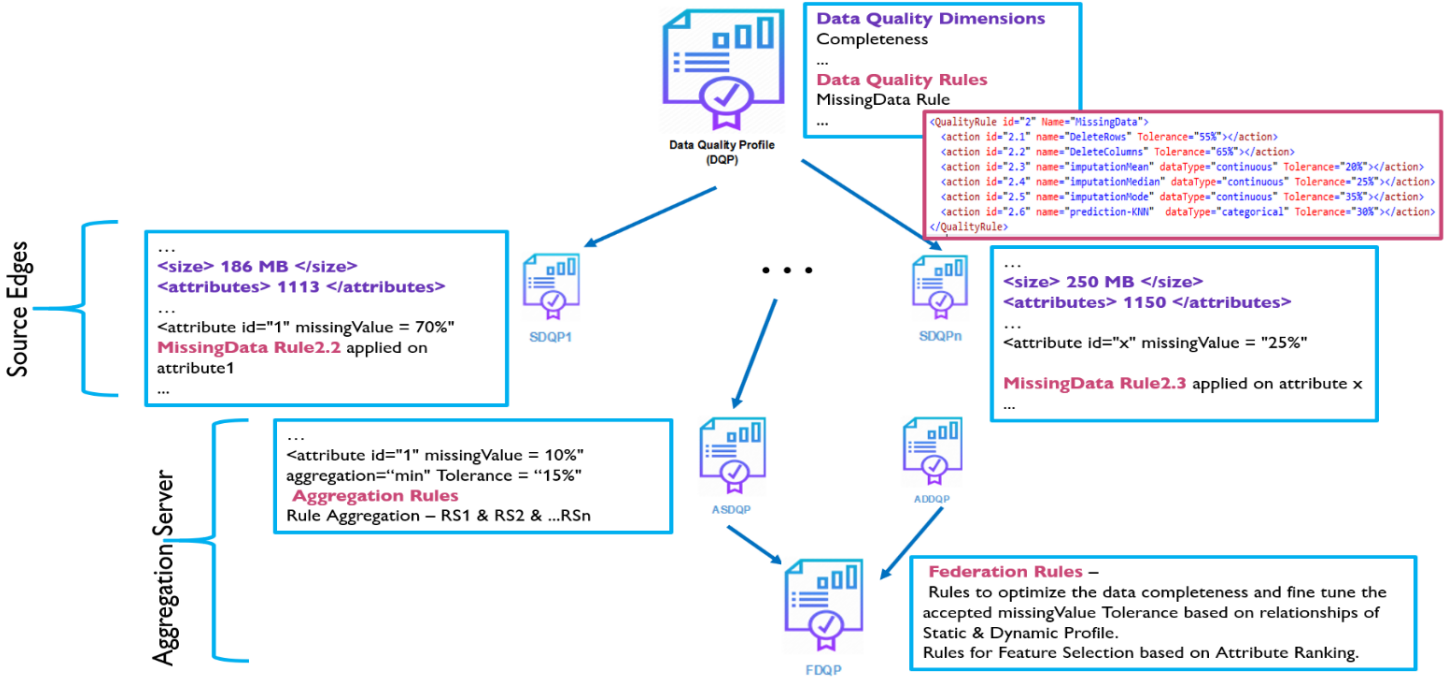


Figure 13: FDQ Profiling Illustrative Example

3.3.4 FDQP Formulation

Data quality aware ($DQ_{FedProf}$), deals with the group aggregation of the DQP received from all the edges. The aggregation is based on the dataset and the data quality dimensions specified in the DQP.

$$DQP_{Fed} = DQ_{FedProf}(Grp\sum(DQP_e(A,W,D,T,R,M))) \tag{3.12}$$

$DQ_{FedProf}$, also considers the dimensionality reduction with feature selection where the number of attributes is reduced based on missing data and acceptable tolerance rules. Dimensionality reduction approaches convert a dataset with dimensionality D into a new dataset with dimensionality d, while keeping as much of the data's shape as possible. For each edge i, and each attribute aj, a missing value MVi information vector is created, where ATol is the acceptable tolerance for the attribute.

$$MV_i = ((a_j.isNull.count()))/D \tag{3.13}$$

$$MV_i a^n_{i,j} = \begin{cases} 1, if = MV_i a_{i,j} < ATol_{(a_{i,j})} \\ 0, Otherwise \end{cases} \tag{3.14}$$

Further, the vector is aggregated during $DQ_{FedProf}$, considering all the edges where rules are applied based on attribute weight and tolerance. For instance, If the attribute weight is above 0.5, the aggregation rule adds a chosen imputation method, however, if the calculated MVi is 0 (meaning it is above the missing value tolerance) and the attribute weight is insignificant, rules are

prescribed to remove the attribute. As a result, these rules override the profile and reduce the dimension to d at the FedProf aggregation. This federated DQP, DQP_{Fed}, is distributed to the edges, where the rules are applied. All of the data measurements and rules acquired in the preceding profiling's are considered in the aggregation. DQP_{Fed} will differ from the *DQPe* in many aspects, including values, size, dimensions, and data metric scores. Federated DQP_{Fed} will be used to determine the right quality metric functions to assess a data quality dimension dk for an attribute ai with a weight wj. Edges that fail to meet critical data quality metrics will be dropped from further processing. Finally, the DQP_{Fed} is sent to the edge nodes and applied for PSN calculation and evaluation based on different PSN fusion algorithms.

3.3.5 Federated Feature Selection

The FDQP relies heavily on federated feature selection to boost classifier precision. The task of feature ranking [245] is to determine the relative importance of a set of features and then ordering them accordingly. The pseudocode for Federated Feature Selection is provided in Algorithm 4. It ranks features according to criteria including feature value, outlier percentage, and missing data percentage; these metrics collectively stand as indicators of data quality. In a real-world scenario, each node does not have the full dataset, making it impossible to accurately compute the importance given to each feature, which is where the federation of feature selection comes in.

The Federated Feature Selection is illustrated in Table 11. Column 1 contains the list of features, while columns 2 through 4 contain the aggregated Feature Selection rank, feature outlier rank, and Feature missing rank. The ranking criteria for each column are specified in the header, for example, the

feature with the lowest outlier percentage is ranked first (i.e., Rank 1), and the feature with the highest outlier percentage is ranked N, where N is the number of features.

The final column Federated Feature Rank is calculated as described in Algorithm 4. In the provided features in Table 11 (A, B, C, D, E), D is the best valuable feature (with the least federated rank) while E is the worst one. Priority is given for feature selection if two features have the same federated value, followed by outlier and missing rank. Thus, missing data and outliers have a negative effect on aggregation for making the right decision regarding data selection and feature extraction, as illustrated by the federated feature selection.

Table 11: Illustration of Federated Feature Selection

Federated Feature Selection				
Ranking Criteria	Best Rank: 1 (Most Feature Value) Worst Rank: N	Best Rank: 1 (least outlier %) Worst Rank: N	Best Rank: 1 (least missing %) Worst Rank: N	Best Rank: 1 (Most Valuable) Worst Rank: N
Feature (N)	**Aggregated Feature Selection Rank**	**Aggregated Feature Outlier Rank**	**Aggregated Feature Missing Rank**	**Federated Feature Rank**
A	3	1	8	12
B	5	4	7	16
C	9	8	2	19
D	1	2	5	8
E	7	23	13	43
...	...	...	...	...

Algorithm 4 Federated Feature Selection Algorithm

Input:
S_n ▷ Participating Edge Source Nodes
DQP_{Features} ▷ Node Features extracted from DQP
$Feature_{\text{Tol}}$ ▷ Tolerance of number of selected features
Output:
$Features_{\text{Fed}}$ ▷ Federated Selected Features

1: **Aggregating Features based on FeatureValue, Outlier and Missing Data from n nodes**
2: **procedure** FEDERATEDFEATURESELECTION(S_n, DQP_{Features}, $Feature_{\text{Tol}}$)
3: $Features_{\text{Agg}} \leftarrow \sum_{e=1}^{n} DQP_{\text{Features}}$(FeatureValue, Outlier, Missing)
4: $FeatureValue_{\text{Rank}} \leftarrow$ sort DESC $Features_{\text{Agg}}$(FeatureValue)
5: $OutlierData_{\text{Rank}} \leftarrow$ sort $Features_{\text{Agg}}$(Outlier)
6: $MissingData_{\text{Rank}} \leftarrow$ sort $Features_{\text{Agg}}$(Missing)
7: $[Feature]_{\text{Rank}} \leftarrow FeatureValue_{\text{Rank}} + OutlierData_{\text{Rank}} + MissingData_{\text{Rank}}$
8: $Features_{\text{Fed}} \leftarrow$ sort $[Feature]_{\text{Rank}}$ limit by $Feature_{\text{Tol}}$
9: **return** $Features_{\text{Fed}}$

Redrafted from: A. N. Navaz, H. T. El Kassabi, M. A. Serhani, and E. S. Barka, "Resource-aware Federated Hybrid Profiling for Edge Node Selection in Federated Patient Similarity Network," MDPI Appl. Sci., vol. 13, no. 23, p. 13114, 2023, doi: https://doi.org/10.3390/app132413114.

This research work is based in [full or part] on the previously published article(s) listed above. I have permission from my co-authors/publishers to use the work listed above in my thesis/dissertation.

3.4 Resource Aware Federated profiling Model

In this section we provide an overview of our resource aware federated profiling model, the formulation of the edge workload profiling and an explanatory example of the edge workload profile.In order to make the most effective use of the distributed computing resources at the edges, the resource aware federated profiling framework was established.

3.4.1 Federated Workload Profiling Model

The federated workload profiling model we propose in this section addresses resource planning. We formulate the planning of edge workload as a resource optimization problem that eliminates the nodes that do not meet the computing demands. The model as shown in Figure 14, is a hybrid resource profiling model that combines both static and dynamic profiling. Static profiling mainly depends on the resource's physical characteristics of the edge, such as CPU computation power, memory size, and transmission power. On the other hand, dynamic profiling relies on the collected real-time available resources and performance metrics at the edge. Real-time available resources include but not limited to CPU utilization, memory usage, and network delays. Whereas,

edge performance metrics are based on data training performance, such as accuracy and convergence time of the model.

Initially, a baseline resource profile is created based on the static profiling and is collected from all edges. An Edge Workload Profile (EWP) is created based on the configuration file available at the edge and a real-time measurement of parameters including the task, incoming data, outgoing data, infrastructure (CPU/GPU, storage), mobility of the edge device, round-trip latency based on current location and the reputation of the edge. The Edge Workload Profiling Module (EWPM), which assumes the responsibility of updating the EWP continuously, receives information about the dynamic characteristics of edge resources from the Node Resource Monitoring Agent (NRMA) at the edge node. Each client creates their own edge workload profile, which is then sent to the federation server, where the Federated Resource Profile is built in accordance with the federation rules. The EWP resource parameters and the edge node's reputation, sustaining prior edge performance, both are taken into consideration by the federation rules. The Federated Workload Profiling (FWP) Module aggregates EWP from all nodes and provides inputs to the Node Selection Module (NSM) that selects the edges depending on resource parameter thresholds that have been stated in advance. The client edge is discarded if the edge resources are limited. For the duration of the data streaming process, iterations are performed, and the profile is incrementally updated. To minimize overhead, only the profile updates are transmitted following the initial run and at every federated aggregation event.

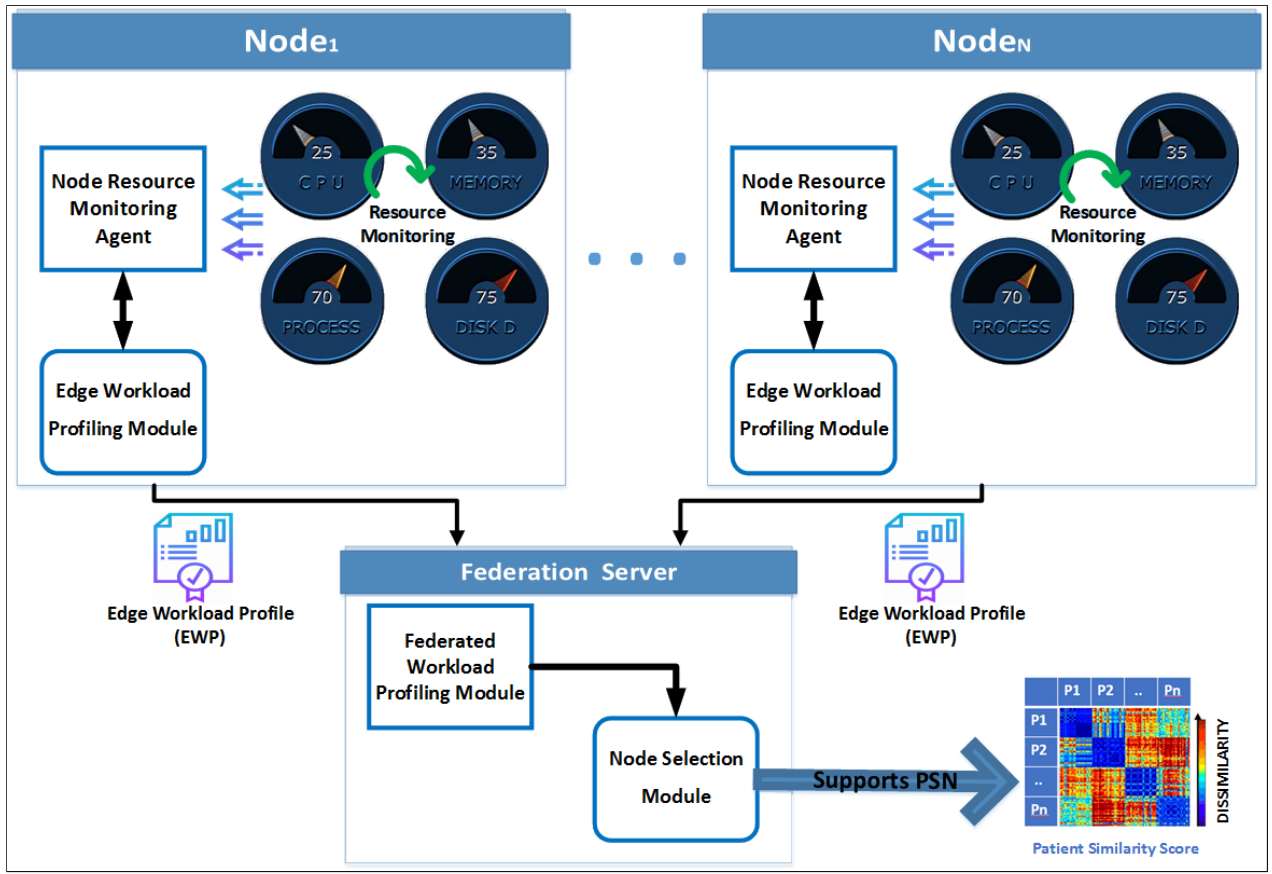


Figure 14: Federated Workload Profiling Model

As time is often a deciding factor in PSN deployments, the objective of this process is to select the nodes with the fastest response times by facilitating faster convergence of the ML algorithms with resource optimizations.

3.5 Research Approach Summary

Our proposed Multi-model PSN Approach to enable Smart Health follows a sequential data value chain enhancement methodology to empower PSN. This methodology began with SCH literature review, specifically precision medicine, and case studies on SCH with a proposed architecture for SCH. Initially we investigated the Multidimensional PSN Data Fusion Model, that aims to build patient similarity from heterogeneous health data sources. This model takes into account various factors that contribute to patient similarity, including demographics, diagnosis, treatment, and clinical outcomes.

To enhance the accuracy of patient similarity prediction, the Data Quality Aware PSN Model was developed. This model ensures data quality at every step along the data value chain with data profiling, federated data profiling, and federated feature selection concepts.

Finally, the PSN Resource Optimization Model was developed to optimize resources through federated resource profiling. This model aids in edge node selection by profiling the available resources and determining which nodes are best suited for specific tasks. In conclusion, Our proposed Multi-model PSN framework ensure accurate patient similarity prediction and efficient resource utilization, enabling effective personalized healthcare delivery.

Chapter 4: Research Evaluation

In the previous chapters, we introduced and discussed three pivotal models that have the potential to reshape the landscape of healthcare and resource management: the PSN Fusion Model, the FPSN Data Quality Aware Model, and the Resource Aware PSN Model. These models represent the culmination of our research efforts, each addressing distinct aspects of data representation, data quality, and resource optimization. We conduct experiments to assess accuracy, data quality, and resource optimization before and after applying these models.

This chapter provides a detailed view of our experimental process, including setup, dataset choices, scenarios, and results.

4.1 PSN Fusion Model - Experimental Evaluation

In this section, we present the experiments conducted to assess the effectiveness of our PSN Fusion Model. We provide details on our experimental setup, dataset choices, and the criteria used for evaluation. Finally, we detail into the specifics of our three experimental scenarios and their corresponding results, offering comprehensive interpretations.

4.1.1 Experimentation Setup

Our experiments were conducted using Google Colab notebooks and a range of machine learning packages, including Scikit-learn, SciPy and TensorFlow. Additionally, we employed BioBERT and BERT with specific configuration details such as attention probabilities, dropout probabilities, and hidden layer sizes, both of which are transformer-based machine learning

models commonly used for NLP pretraining in our batch processing. To assess the performance of the matrices, we carried out evaluations using JAVA within the Apache NetBeans IDE version 12.2.

4.1.2 Dataset

In our experiments, we utilized two distinct datasets:

Dataset-1 - Covid-19 Epidemiological Data: This dataset, sourced from epidemiological Covid-19 data [246], was meticulously compiled from state, regional, and local health reports. It is geocoded and encompasses various patient information, including symptoms, primary dates (onset, admission, confirmation), chronic diseases, travel history, and admission status. Our analysis focused on data up to August 30, 2020, resulting in 155 complete and preprocessed patient records. Each record represents an individual patient case and comprises 33 columns. These columns correspond to four class outcomes: death, discharged, stable, and recovered. This dataset was chosen for experimenting with clinical text data, particularly for the application of Natural Language Processing (NLP).

Dataset-2 - Framingham Offspring Heart Study: The Framingham offspring heart study [247] is a long-term cardiovascular cohort study involving adult offspring of the original Framingham study, which commenced in 1949 in Framingham, Massachusetts, USA. The study recruited a total of 5,124 individuals from 1971 to 1975 and tracked them over the years to analyze trends in cardiovascular disease and its risk factors. Additionally, it aimed to investigate the correlation between risk factors and the incidence of

cardiovascular diseases such as stroke, myocardial infarction, and cardiovascular disease-related mortality.

Table 12: Summary of the Datasets Used in the Experiments

	Dataset-1	**Dataset-2**
Data based On	Covid-19	Cardiovascular disease (CVD)
Type	Static	Static and Dynamic
Size	Small (200)	Big (20000)
Source	Curated data by Scientific Data curation team \[246]	Framingham offspring heart study \[247]
Fields	Static: ID, age, gender date_onset_symptoms, date_admission_hospital, date_confirmation, additional_information, chronic_disease_binary, chronic_disease, symptoms, outcome	Static: PID, exam_age, gender, smoke, diab, hypermed, age_baseline, smoke_baseline,gender_baseline, diab_baseline, hypermed_baseline, time_long_years, time_to_event_years Dynamic: Bmi, sbp, dbp, chol, hdl, ldl, trig non_hdl, chol_hdl_ratio, time_long_years, time_to_event_years, time_long_scal, time_to_event_scal

We chose this dataset for our experimentation because it captures dynamic patient data characteristics and includes multiple visit records and static features for each patient, making it ideal for evaluating our fusion algorithm. Table 12 summarizes the key features of the two datasets used in our experiments.

4.1.3 Evaluation Criteria

In personalized prediction, patient similarity plays a crucial role. To assess patient similarity, we employ KNN-similarity, where 'k' represents the number of patients and is determined through distance measurements. However, our primary objective is to gauge similarity based on patient outcome categorization, for which we use a supervised learning approach.

In our similarity evaluation, we forecast if each pair of patients will be diagnosed with the same outcomes or not, and the distance similarity measure is within tolerance or a threshold. More specifically, if two patients have the same label information and are near in proximity, they are considered similar.

Our experiments involve comparing various distance algorithms to generate our proposed similarity matrices, aiming to identify the optimal one. Additionally, we assess the performance of the fused matrix in comparison to the static and dynamic data similarity matrices independently. Our evaluation employs diverse criteria, including accuracy, precision, recall, and F1-score [248], to measure the effectiveness of these similarity matrices.

A two-by-two confusion matrix depicting all four possible outcomes sums up our similarity matrix model evaluation: true positive (TP), false positive (FP), false negative (FN), and true negative (TN).

TP: accurate prediction of similar patients (predicted that two patients are similar).

TN: accurate prediction of non-similar patients (predicted that two patients are not similar and both have different outcomes. e.g., P1 died and P2 survived).

FN: similar patients inaccurately predicted as non-similar (predicted two patients as non-similar but they both have similar outcomes).

FP: non-similar patients inaccurately predicted as similar patients (predicted two patients as similar but they have different outcomes).

We have adopted the aforementioned measurements to validate and compare the performances of our similarity matrices as follows.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{4.1}$$

$$= \frac{(correctly\ predicted\ similar\ and\ non-similar\ patients)}{(total\ number\ of\ predictions)}$$

$$Recall = \frac{(TP)}{(TP+FN)} \tag{4.2}$$

$$= \frac{(correctly\ predicted\ similar\ patients)}{(correctly\ predicted\ similar\ patients + similar\ patients\ incorrectly\ predicted\ as\ non-similar)}$$

$$Precision = \frac{(TP)}{(TP+FP)} \tag{4.3}$$

$$= \frac{(correctly\ predicted\ similar\ patients)}{(correctly\ predicted\ similar\ patients + non-similar\ patients\ incorrectly\ predicted\ as\ similar)}$$

$$F1Score = \frac{2*(Recall*Precision)}{(Recall+Precision)} \tag{4.4}$$

4.1.4 Deep Learning Configurations

In this subsection, we detail the configurations of the DL models employed for our research, specifically focusing on the BERT model for NLP tasks and the LSTM Autoencoder for dimensionality reduction.

In our experiments, we utilized the BERT model for NLP tasks. The BERT configuration specifics are instrumental for achieving reproducibility and understanding its performance dynamics. Table 13 outlines these configuration details.

Table 13: Configuration Details for BERT

Parameter	**BERT**
Attention Probabilities Dropout	0.1
Activation Function	GELU
Hidden Dropout	0.1
Hidden Size	768
Initializer Range	0.02
Intermediate Size	3072
Max Position Embeddings	512
Number of Attention Heads	12
Number of Hidden Layers	12
Type Vocabulary Size	2
Vocabulary Size	28,996

In the BERT configuration, essential parameters like the Activation Function, represented by the Gaussian Error Linear Unit (GELU), and the Number of Attention Heads shape the model's behavior. The "Attention Probabilities Dropout" and "Hidden Dropout" manage overfitting, the "Hidden Size" and "Number of Hidden Layers" determine its complexity, and the "Max Position Embeddings" and "Vocabulary Size" establish sequence constraints and unique tokens, respectively.

Subsequently, for sequence data processing and dimensionality reduction, we adopted an LSTM Autoencoder. The configuration specifics of the LSTM Autoencoder play a key role in determining its operational attributes. Table 14 showcases these configurations.

In the LSTM Autoencoder configuration, the "Input Size" and "Hidden Size" denote the number of expected features and features in the model's hidden state, respectively. The model's depth, indicated by the "Number of Recurrent Layers", determines how many LSTM layers are stacked, while "Sequence

Length" adjusts based on patient visits, and the tensors h_0 and c_0 represent initial hidden and cell states, respectively, for the LSTM layers.

Table 14: Configuration Details for LSTM Autoencoder

Parameter	**Value**
Input Size (features)	9
Hidden Size	32
Number of Recurrent Layers	1
Batch Size	32
Sequence Length	Variable (depending on patient visits)
h_0 Tensor Shape	[1, 32, 32]
c_0 Tensor Shape	[1, 32, 32]

The choice of these configurations was guided by the dataset's needs, research challenges, and main goals, enabling effective model development and assessment.

4.1.5 Objective 2: Experimentation

We conducted a series of experiments to assess our multidimensional Patient Similarity Network (PSN) using two distinct datasets, focusing on two primary experimental scenarios, and an additional benchmarking scenario.

Scenario 1: In this scenario, we aimed to evaluate the PSN model using data containing static features and a mix of numerical and textual clinical data. Our objectives included:

- Predicting ICU admissions for Covid-19 patients based on Dataset-1.
- Assessing the accuracy of the patient similarity matrix by employing NLP models like BERT and one-hot-encoding to capture clinical textual data semantics.

- Identifying the optimal similarity distance measurement method among Euclidean, Manhattan, cosine, Chebyshev, and weighted Manhattan approaches.
- Determining the ideal weight distribution among features when using the weighted distance evaluation approach to improve accuracy.
- Evaluating the PSN model's performance when applying the local similarity approach to enhance accuracy.

Scenario 2: In this scenario, we aimed to evaluate the overall performance of our multidimensional model using a dataset containing both dynamic and static features. Our objectives included:

- Predicting future Cardiovascular Disease (CVD) events based on Dataset-2.
- Constructing a static PSN matrix for the static portion of the data and evaluating the STPS matrix's performance.
- Assessing the performance of an autoencoder used for dynamic patient data to achieve data reduction and represent the information in a lower-dimensional space.
- Constructing and evaluating the dynamic similarity matrix.
- Evaluating the performance of the fused similarity matrix based on our SNF algorithm, verifying its capability to represent large, heterogeneous, and dynamic dataset contents.

Scenario 3 - In this scenario, we extend our evaluation to benchmark our multidimensional PSN model against other classification algorithms commonly used in predictive modeling.

Throughout our experiments, we compared different geometric distance algorithms (e.g., Euclidean, Manhattan, cosine, Chebyshev, and weighted Manhattan) for patient similarity calculations. These scenarios align with our goals to validate the PSN model's effectiveness in various data contexts.

4.1.5.1 Scenario 1. PSN Evaluation on Static Data having Numerical and Textual Data

Dataset-1 is used for this scenario, wherein both numerical and clinical textual data are available. The effectiveness of the static algorithm solution and distance estimation are evaluated. Further, the classification performance is analyzed using a fivefold cross-validation method. The accuracy, recall, precision, and F1-score measures are calculated, as explained in the evaluation criteria of this study, to compare the performances of different similarity distance calculation algorithms.

A. Accuracy Measure of Patient Similarity

In this scenario, we generate numerical representations from the contextual embedding of textual clinical data via hot encoding and BERT. Next, we evaluated the accuracy of the resulting patient similarity matrix using different distance calculation techniques, including Euclidean, Manhattan, cosine, Chebyshev, and weighted approaches (Figure 15). The graphs illustrate that the Euclidean and weighted distance calculations performed better in accuracy for one-hot encoding, whereas cosine excelled when using BERT.

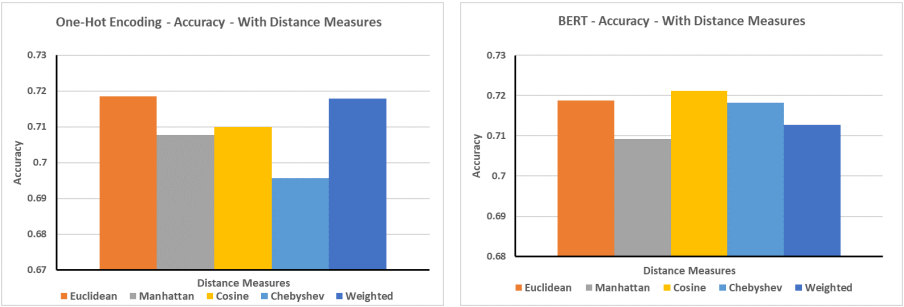


Figure 15: Accuracy with Various Distance Measures (One-hot Encoding and BERT)

Table 15: Evaluation of the PSN Distance Measures with One-hot Encoding and BERT

	One-Hot Encoding			**BERT**		
	Accuracy	**Precision**	**F1-score**	**Accuracy**	**Precision**	**F1-score**
Euclidean	0.71856	0.721006	0.83348	0.72365	0.99729	0.83725
Manhattan	0.70776	0.710064	0.826226	0.72275	0.99891	0.83703
Cosine	0.71	0.712431	0.826848	0.84602	0.97635	0.89971
Chebyshev	0.69576	0.71904	0.80984	0.72122	0.9966	0.83591
Weighted	0.71792	0.723265	0.828192	0.71827	0.96927	0.83037

Table 15 presents the results obtained based on the performance evaluation parameters of various distance measures used in one-hot encoding and BERT. The overall performance of BERT is slightly better than that of one-hot encoding.

B. Weighted-Distance Accuracy Measure Against Similar Patients

In this experiment, we evaluated the patient similarity matrix generated using the weighted Manhattan distance algorithm after BERT contextual encoding. We defined different weights for each feature to provide more significance to some features over others that was validated based on medical expertise.

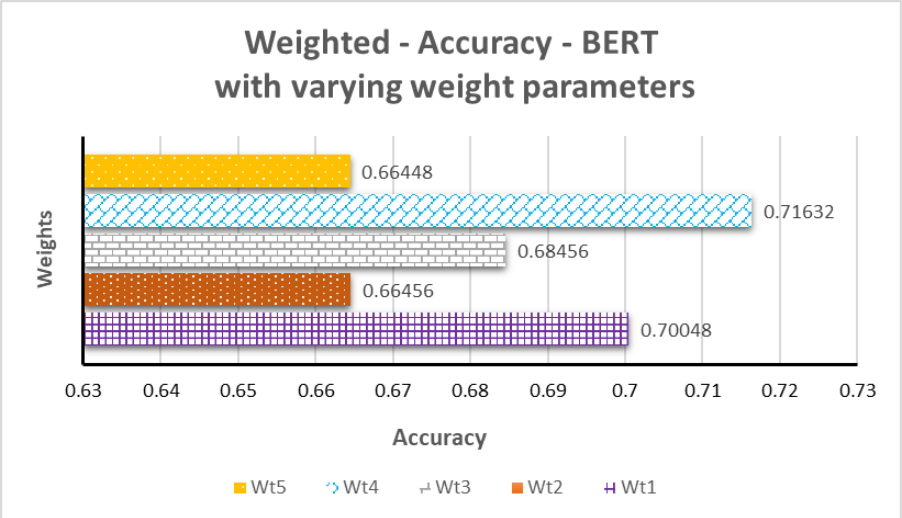


Figure 16: Weighted Accuracy Based on Weighted Features

Further, we have employed weighted scoring [249], a prioritization framework to prioritize by scoring the features and determined the weights for our experiments. Wt1 = [1,1,3,3,3,1], Wt2 = [1,1,3,2,3,3], Wt3 = [1,1,4,3,2,2], Wt4 = [1,1,3,3,3,3], and Wt5 = [1,1,3,1,1,1] represent the sets of weights given to age, gender, symptoms, additional_information, chronic_disease and chronic_disease_binary, respectively. Optimal results were obtained when the features (symptoms, additional_information, chronic_disease_binary, and chronic_disease) of Wt4 were given higher weights, as depicted in Figure 16, which highlights the importance of feature selection.

C. Accuracy Measure against the Selected Percentage of Similar Patients

Our next step in the experiment is based on the strategy of using the K-nearest neighbors of similar patients to calculate the local similarity for each matrix to increase the prediction accuracy. The details of this approach are depicted in this study.

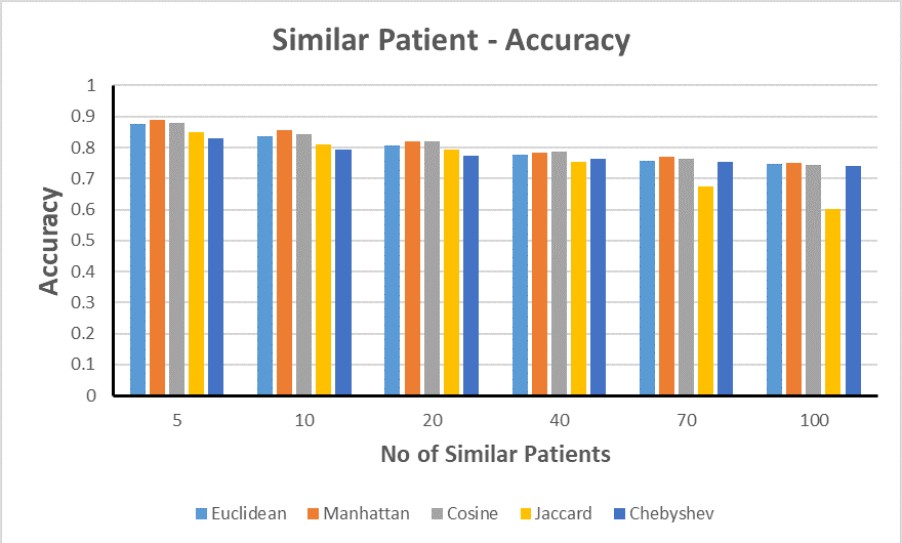


Figure 17: Accuracy with Varying Training Data Involving Similar Patients

The results presented (Figure 17), show that improved outcome prediction results can be obtained by considering only similar patients. The highest accuracy of 89% could be obtained for the Manhattan approach when selecting 5% of the related patients in our training, whereas selecting the full data (100%) resulted in a mere 75% accuracy. Thus, selecting the optimum number of similar patients is crucial to improve the predictive performance and decrease the training time (a key factor when big health data are considered).

4.1.5.2 Scenario 2. Multidimensional PSN Data Fusion Model Evaluation with Static and Dynamic Features

In this scenario, we adopt Dataset-2, which is a combination of patient static demographic and dynamic longitudinal data, indicating multiple patient visits, which is ideal for evaluating our proposed fusion model. The class attribute in this dataset indicates the possibility of developing CVD.

A. Static PSN Evaluation

In this experiment, we evaluate the accuracy of the STPS matrix based on different distance calculation algorithms. Table 12 presents the static features used for similarity. We evaluated the accuracy based on different K-nearest neighbor values where K represents the % of similar patients. The accuracy increased when closely similar patients were selected for training the model, that is, the K-value decreased as depicted (Figure 18). The Jaccard distance resulted in the highest accuracy of 96% among all trials when considering 5% similar patients and 90% when using the full dataset. All the remaining distance measures resulted in improved results when the training data included most similar patients.

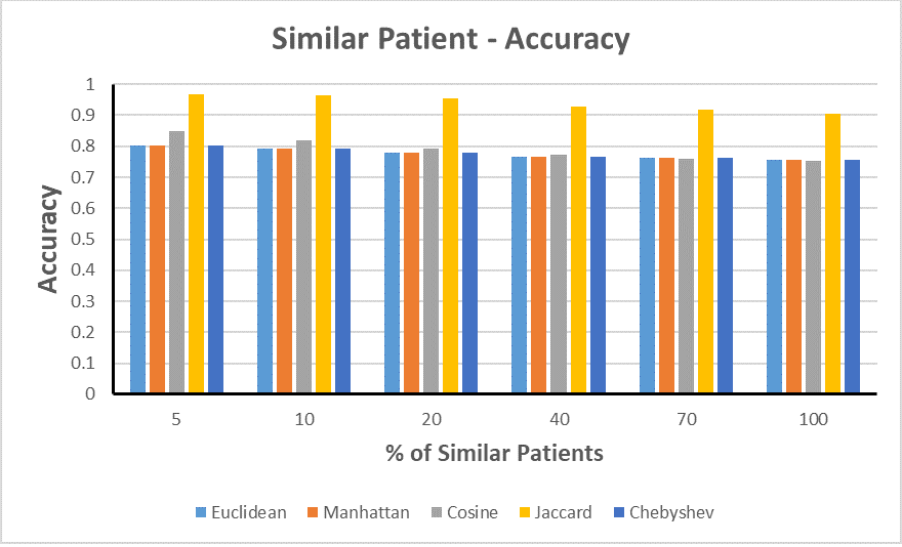


Figure 18: Static Data: Accuracy in Case of Similar Patients

B. Dynamic PSN Evaluation (Autoencoder)

In our Dataset-2 (CVD dataset), each patient has a different number of records representing the health measurements associated with each visit, which dictates reducing data dimensionality to facilitate the construction of the dynamic PSN. We first train our dataset utilizing a reconstruction autoencoder model to reduce the size from 20680 to 4046 rows with 5D embeddings each. Subsequently, we train the autoencoder model-generated output into a similarity matrix using one of the different distance measurement approaches. First, we split our dataset into static profile data and dynamic time-series patient visit records.

Figure 19 presents the dynamic data balanced distribution, for example, approximately 400 patients have two records each, and 800 patients have seven records each.

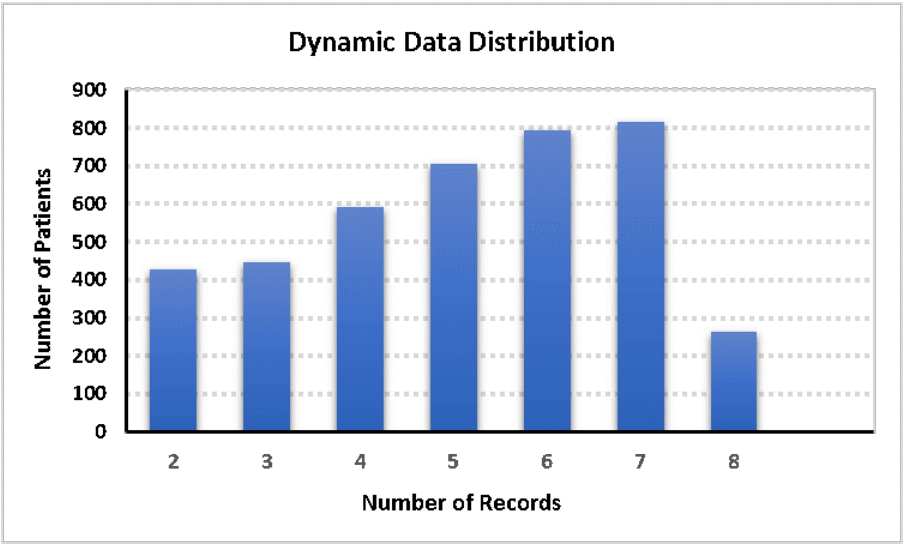


Figure 19: Dataset 2: Dynamic Data Distribution

This dynamic part of our data is fed into an LSTM layer. The proposed model takes a batch of series of patient exam records as input and outputs (1 × 5) vector that is the final hidden state. However, the decoder uses the (1 × 5) vector and passes it to an LSTM layer, which produces the dynamic time-series part. Figure 20 shows the architecture of the encoder-decoder neural network developed for data reduction.

Figure 21 illustrates the autoencoder reconstruction loss values obtained based on Mean Square Error (MSE) while generating (1 × 5) vector embedding. In this model, the reconstruction loss values decreased gradually and stabilized after approximately 25000 iterations.

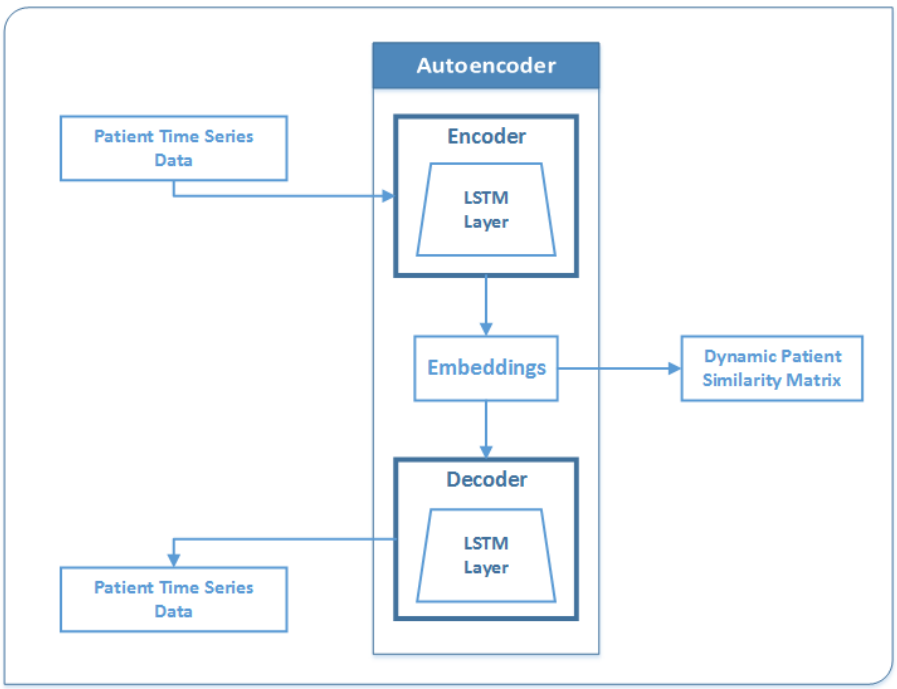


Figure 20: The Architecture of Data Reduction Autoencoder

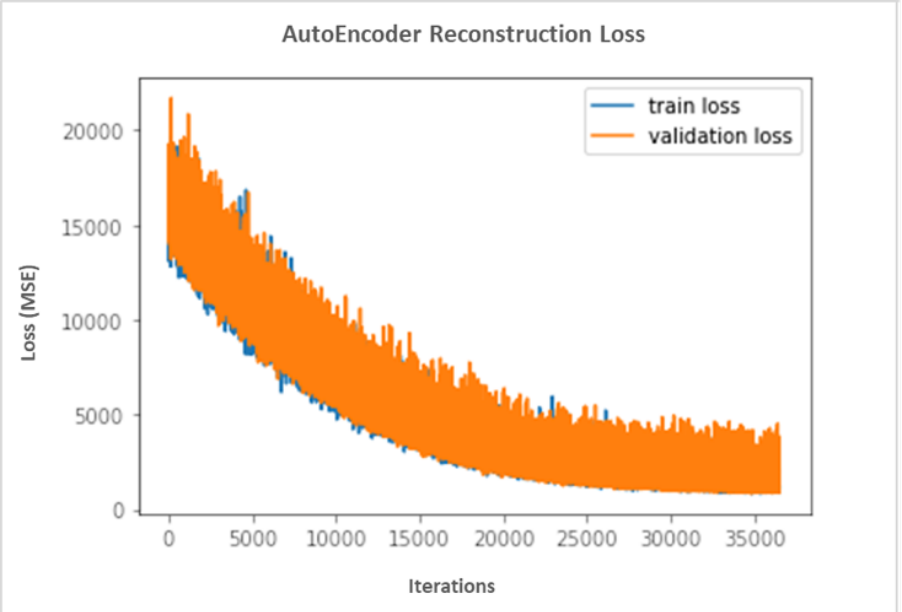


Figure 21: Reconstruction Loss Associated with an Autoencoder

Choosing LSTM in our Autoencoder model facilitates the feature reduction process to learn from the temporal relationships among time-series features, instead of implementing a feature reduction process that flattens all the time-series features and loses the temporal information contained in the set of features and Choosing 5D embeddings produced good accuracy results when training.

C. Multidimensional PSN Data Fusion Model Evaluation

In this experiment, we evaluated the performance of the resultant fused patient similarity matrix against the outcome class with respect to the different distance measurements explained in this study.

Figure 22 depicts the performance of the final PSN matrix when compared with the static and dynamic similarity matrices while adopting

different distance measurements. Our proposed SNF approach improved the accuracy of the final fusion patient similarity matrix when compared with the accuracies of the static and dynamic similarity matrices.

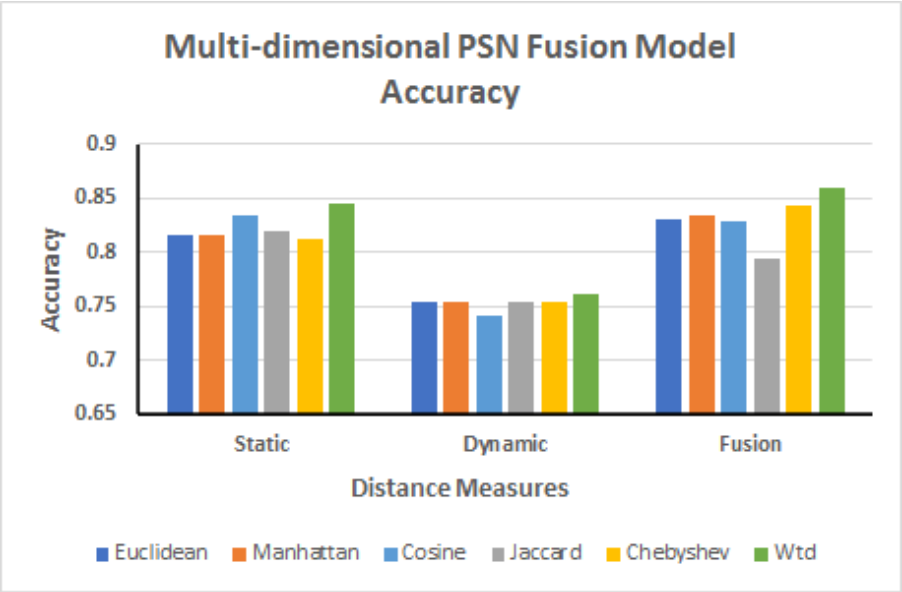


Figure 22: Accuracy- Multidimensional PSN Data Fusion Model VS Static & Dynamic PSN

Our experimental evaluation (Figure 22), also discloses that the static PSN data provided more accuracy than the dynamic PSN data. Here the dataset consisted of static data such as gender, age, diabetic status, etc. which featured categorical values with little variance. However, the dynamic features included frequently changing time-variant fields such as BMI, Chol, LDL, etc., and each patient had a varying number of hospital visits (Figure 19). According to our view, the variance in static and dynamic data components, as well as the differences in PSN calculation methods, such as data reduction using

autoencoders in dynamic PSN calculation, resulted in a considerable difference in accuracy.

We found a couple of works that studied autoencoder accuracies that validates our experiment. An article that compares autoencoders with PCA [250] confirms the fact that the PCA and autoencoder data reductions behave similarly. However, PCA outperforms Autoencoding and the network itself needs to be tuned for the particular task to do better. A study on the performance of autoencoder with Bi-Directional LSTM [251] reports that the accuracy and F1-score of the model with autoencoder drop by around 4% and 9%, respectively, indicating that some information is lost because the encoding process does not hold all of the information from the original data. As per Chen [252], even if the epoch size is high, the accuracy will be less than the initial accuracy because encoding and decoding cause some data loss. We believe this holds true in our above experiment using autoencoder for data reduction as well, where accuracy variation is around 9% between the static and dynamic PSN data.

4.1.5.3 Scenario 3. Benchmark to Other Classification Algorithms

Our multi-model PSN can be used for unsupervised or supervised data with high accuracy. To validate that, we selected one of the features as a labeled outcome to convert unsupervised learning into a supervised learning technique. Further, we evaluated the similarity network matrices with respect to this outcome. Experimental results show that higher accuracy is achieved by the fused similarity matrix when compared with those of both the static and dynamic data similarity matrices when evaluated independently.

Furthermore, we benchmarked our PSN model with other widely adopted classification algorithms using the CVD and COVID-19 datasets. Higher accuracy can be obtained by performing classification using our multi-model PSN when compared with those of other baseline-supervised classification models such as LR, naïve Bayes, zeroR, decision tree, and random forest. Table 16 presents the different accuracy results of the classification algorithms. When testing using the CVD dataset results, the accuracy improved by 20% when compared with that of naïve Bayes; further, a minimum of 10% improvement could be observed when compared with those of zeroR and decision tree. However, experiments on the COVID-19 dataset show that our model results in a 7% higher accuracy than those of zeroR and LR.

Table 16: Benchmark Multi-dimension PSN Fusion model to Other Classification Algorithms

Accuracy %	**PSN**	**Naïve Bayes**	**zeroR**	**Logistic Regression**	**Random Tree**	**Decision Tree**
COVID-19 Dataset 1	89	84.8	83.2	83.2	88.8	86.4
CVD Dataset 2	96	80.67	87.03	87.10	87.32	87.03

4.1.6 Results Discussion

Our experiments, spanning three critical scenarios, assessed the effectiveness of our multidimensional Patient Similarity Network (PSN) model.

In Scenario 1, different distance measurement approaches impacted accuracy variably. Notably, Euclidean and weighted Manhattan performed well with one-hot encoding, while cosine excelled with BERT. Weighted-distance accuracy improved with feature prioritization, validated by medical expertise.

In Scenario 2, static data outperformed dynamic data, attributed to differences in variance and PSN calculation methods. Autoencoder-based data reduction showed a trade-off between dimensionality reduction and information loss.

Scenario 3 featured benchmarking against common classification algorithms, highlighting our multi-model PSN's competitive edge, particularly in CVD and COVID-19 datasets. The fused similarity matrix consistently outperformed static and dynamic matrices.

In summary, our experiments showcased the versatility and potential of the multidimensional PSN model across diverse data scenarios. It offers enhanced accuracy, especially with static data, and excels in benchmarking against established classification algorithms. These results underscore the utility of our PSN model in advancing predictive healthcare.

4.2 FPSN Data Quality Aware Model - Experimental Evaluation

This section describes the experiments conducted to evaluate our proposed FDQ profiling model. The primary goals are to evaluate the data quality and the accuracy of the training model both before and after the FDQ profiling model is applied to the data at the edges. In the subsections that follow, we will describe the various aspects of the experiments conducted, such as the experimental setup, the dataset employed, the experimental design,

and the various scenarios tested. We will conclude by discussing the results, which depict an improvement in accuracy, and the reasoning behind it. Initial Assumption: We assume that the data collected at the nodes are homogenous and identically distributed (i.i.d).

4.2.1 Dataset

The dataset we chose for our experiments contains 2,126 instances and 23 attributes derived from cardiotocograms, which are continuous measurements of the fetal heart rate using an ultrasound transducer placed on the mother's abdomen and categorized by expert obstetricians. The parameters used for data analysis are instantaneous fetal heart rate (FHR) and simultaneously communicated uterine contraction signals. The classification results were based on the fetal state labels (N = normal; S = suspect; P = pathologic) [253].

For the selection stage and reduction of attributes, in particular the numerical input data and a categorical (class) target variable, there are two well-known feature selection techniques mainly ANOVA-f Statistics and Mutual Information Statistics. The results of this test can be used for feature selection by removing from the dataset those features that are independent of the target variable.

4.2.2 Experiment Setup

All of our experiments for this study were conducted in IPython, an enhanced interactive Python for the experimental configuration that includes multiple machine learning libraries for DL and FL. The entire Fetal Health dataset is randomly divided into five datasets each corresponding to one of the

five edge nodes. The dataset has few quality issues, so we synthesized errors and noise at the dataset's edges to reflect a true clinical distribution and to illustrate how FDQP will improve it.

4.2.3 Objective 3: Experimentation

To evaluate our proposed FDQ profiling model, we implemented eight scenarios. The first scenario evaluates the data quality profiling with the main focus on accuracy considering baseline accuracy, after missing data imputation, applying Local Data Quality Profiling (LDQP) which includes feature selection and rules application. The second scenario illustrates the node selection criteria defined in FDQP, which takes into consideration the completeness and consistency of DQ parameters. In the third scenario, the Federated Feature Selection and ranking are evaluated. Scenario 4 compares the accuracy of the FDQP to that of the LDQP and the baseline. In Scenario 5, the accuracy, completeness, and consistency of the data quality metrics are examined before and after FDQP. Scenario 6 analyzes the accuracy of various classifiers using FQQP to determine the most accurate classifier. The seventh scenario indicates the number of features selected and the associated training time at each node. Finally, scenario 8 depicts the PSN accuracy before and after FDQP.

4.2.3.1 Scenario -1

Create a Data Quality Profile (DQP) based on the XML file containing the data set and sends it to the edge nodes. The quality characteristics of the profile are updated by generating Local Data Quality Profile (LDQProfile) and sent to the Server. As shown in Figure 23, LDQProfile is applied at the edge

node and the experiment's accuracy is evaluated node-by-node with baseline and after each of the processes (Missing Data Imputation, Rows Removed, LDQP).

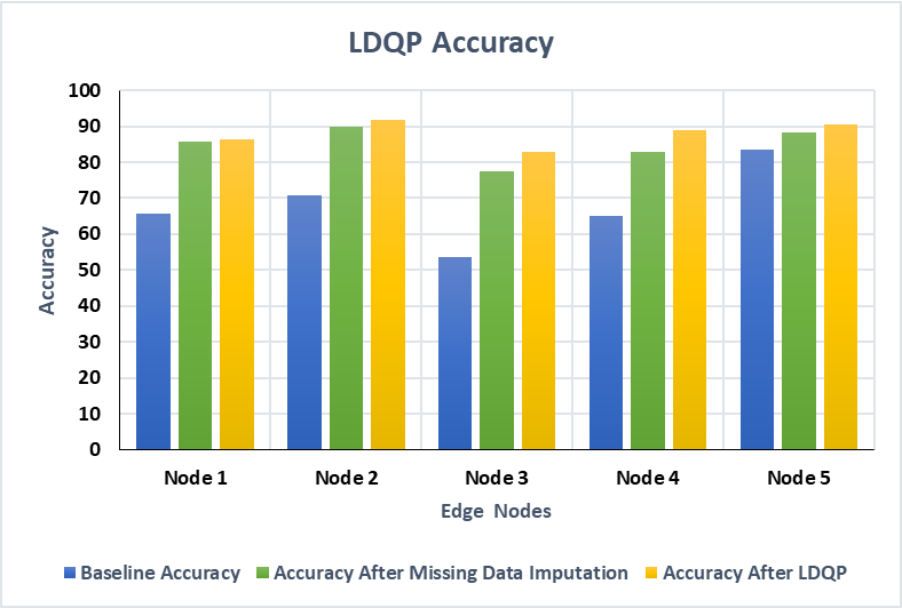


Figure 23: Local Data Quality Profiling (LDQP) Accuracy Evaluation

4.2.3.2 Scenario -2

Edge node selection occurs before FDQP is applied at the edge node. In other words, LDQProfile will be used to select nodes on the server. If the DQ metrics values (Accuracy, Completeness or Consistency) are less than the tolerance, the node is eliminated. One of the first and most important steps in our proposed FDQP is assessing the edge-level distribution of classes and establishing criteria for node selection.

Figure 24 illustrates the problem with the consistency and completeness of the Node-3 dataset. The FDQP criteria determine what levels

of information must be present in a node's representation. Accordingly, Node-3 is eliminated from further processing after failing to comply with the DQ requirements.

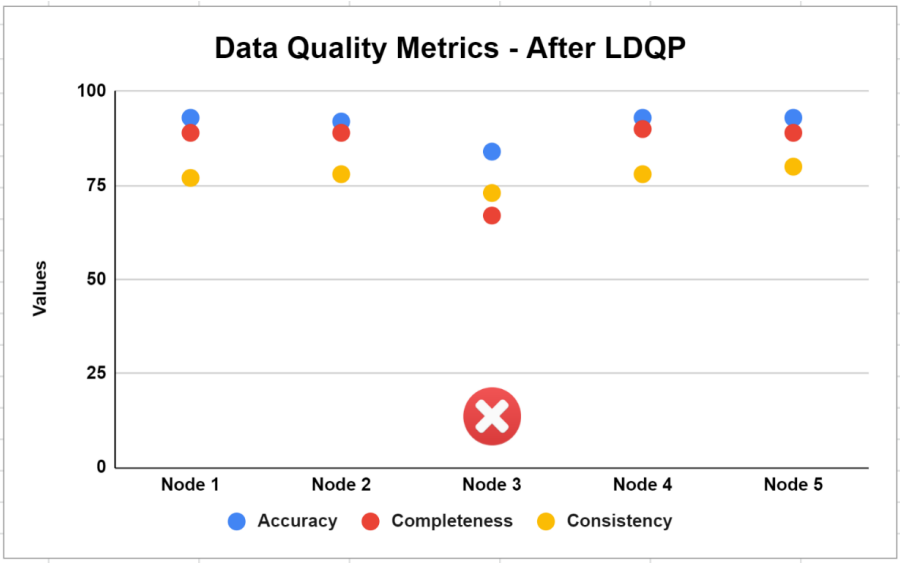


Figure 24: Node Selection Based on LDQProfile Data Quality Metrics

4.2.3.3 Scenario -3

The server federates the profile data from the edges. Features are eliminated (feature selection) according to the feature selection algorithm, and the retained features are documented in the feature driven FDQProfile. The resulting FDQProfile includes data imputation rules and is transmitted to the edge nodes. In order to assess the PSN's accuracy and precision, it is necessary to apply FDQProfile to the edge node. Figure 25 depicts the federated feature rank for each of the attributes in our experimental dataset.

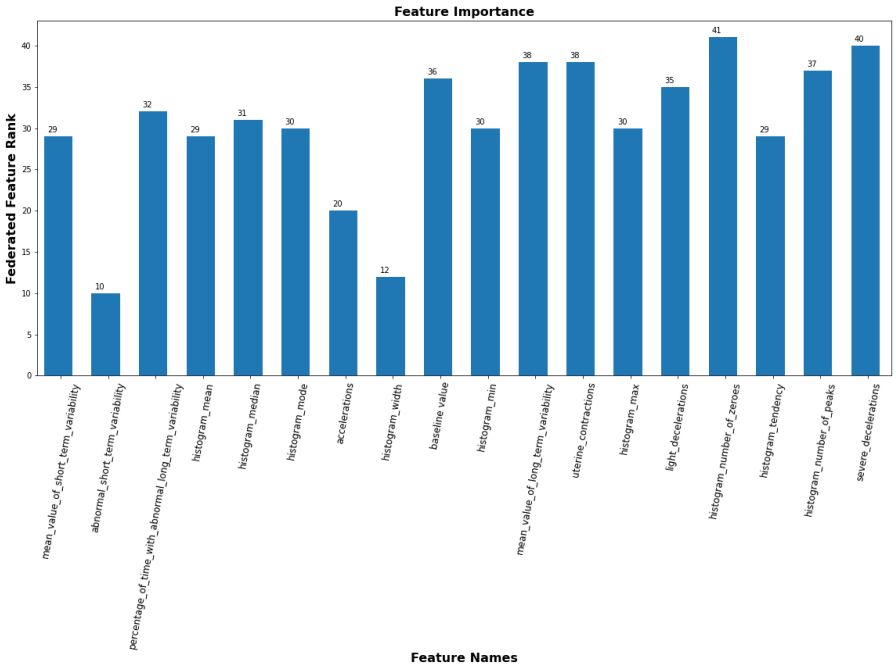


Figure 25: Federated Feature Selection

4.2.3.4 Scenario -4

Evaluation of FDQ profiles is one of the most crucial aspects of our experimentation. Node 3 has already been deleted, and accuracy after FDQP for the remaining nodes is compared to accuracy after LDQP and baseline accuracy as seen in Figure 26. We can see that accuracy has improved significantly around 10% with LDQP and further to a maximum of 5% with FDQP.

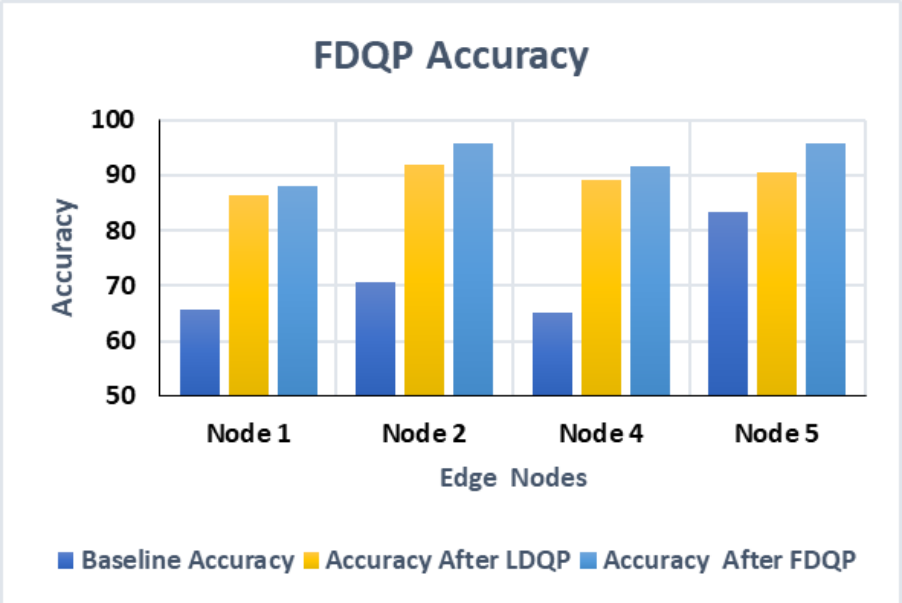


Figure 26: Accuracy (Baseline, After LDQP and After FDQP)

4.2.3.5 Scenario -5

Before and after applying FDQP, we analyzed the metrics for data quality. Figure 27 shows that the coefficient of variation was reduced following FDQP, suggesting enhanced consistency and that the completeness factor was increased to 100%, indicating that certain data quality issues had been addressed by the process.

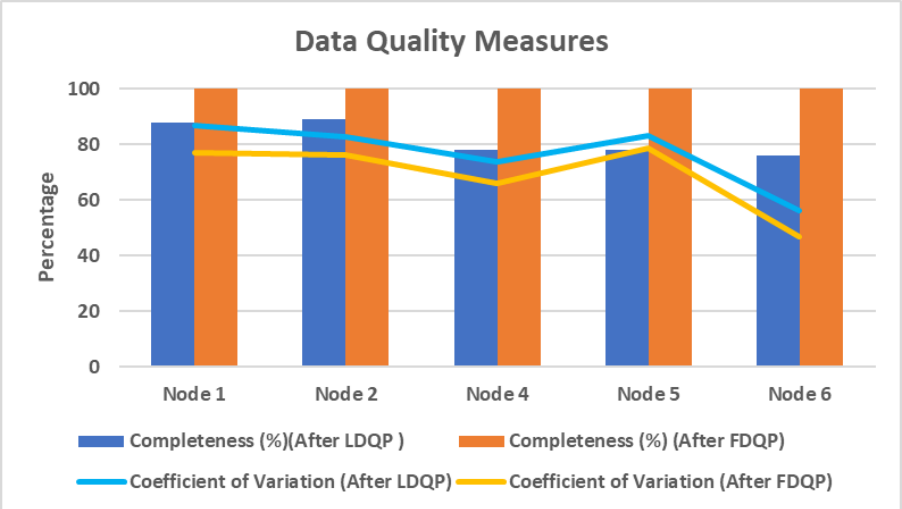


Figure 27: Data Quality Metrics Assessment After LDQP and FDQP

4.2.3.6 Scenario -6

In this scenario, an evaluation is performed using various ML models to determine which classifier is the most accurate in forecasting fetal health at each node, and the results are displayed in Figure 28.

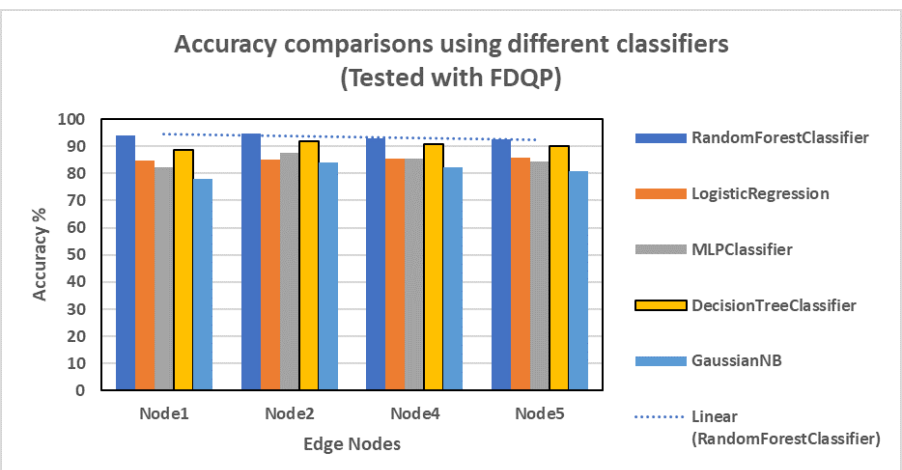


Figure 28: Accuracy Comparisons with Different Classifiers

Both the Random Forest Classifier and the Decision Tree Classifier have been shown to have the best performance in all of the edge nodes. Edge 2 had the greatest accuracy gain, with Random Forest Classifier attaining 95% accuracy.

4.2.3.7 Scenario -7

The selection of features is one of the primary characteristics of our suggested FDQP, and its evaluation can be found in Figure 29. We can see that the number of features has been reduced in each of the nodes, which has led to a shorter amount of training time when compared to the initial amount of training time.

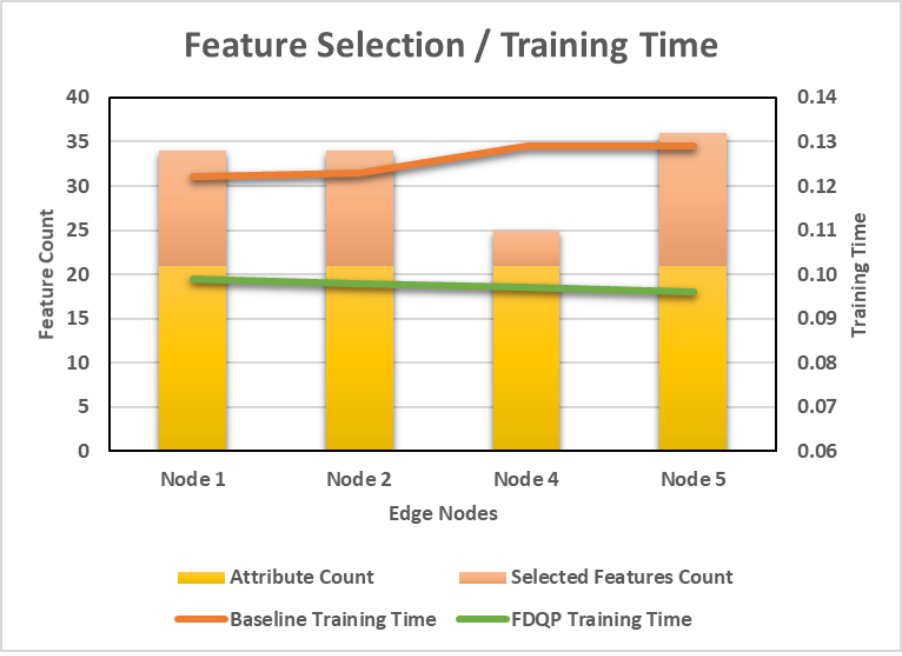


Figure 29: FDQP Feature Selection VS Training Time

4.2.3.8 *Scenario -8*

Patient similarity evaluation is performed after FDQProfile is reviewed, and we observed in Figure 30 that FDQP has unquestionably increased the accuracy, with an average increase of 7% and a maximum gain of 9% accuracy.

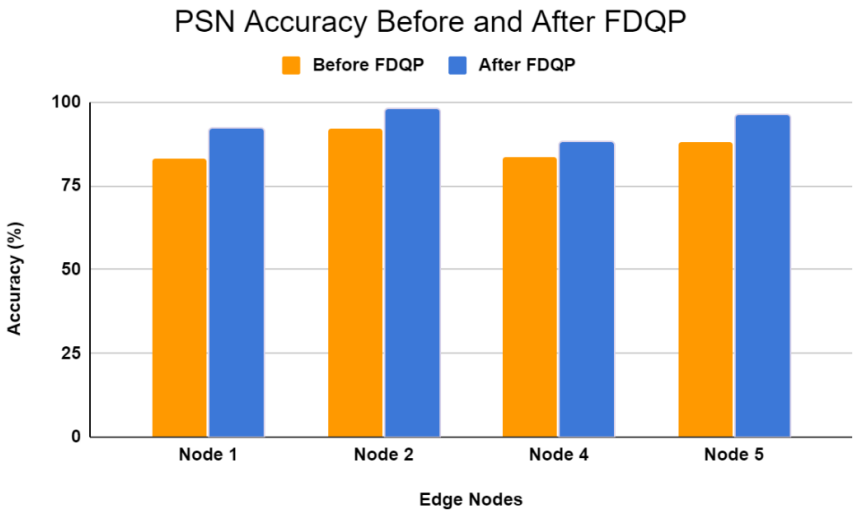


Figure 30: PSN Accuracy Before and After FDQP

4.2.4 *Results Discussion*

Profile aggregation with Federated Feature Selection of attributes from various nodes enhances discriminative feature efficiency and limits ineffective ones. They are calculated by feature aggregation and then optimized via the proposed rules for elimination. The feature with low weight is eliminated in the experiments, which has been found to be effective in improving the accuracy. In addition, we can see that the results obtained by profiling the data at multiple nodes are, in general, more stable with the overall ranking concerning management of resources. Because combining the outputs of

several experts yields better and more robust results than a single expert, the concept of distributing features across nodes and then federating the profile's results into a final one is similar to that of ensemble learning or a mixture of experts. The research also demonstrated the significance of concentrating on DQ components and selecting feature nodes according to data quality metrics.

4.3 Resource Aware FPSN Model - Experimental Evaluation

This section details the experiments devised to evaluate the proposed federated resource profiling model. In the proposed approach, quality profile federation plays a crucial role, and involves aggregating multiple quality profiles, which represent the resource attributes, tolerance values, and rules for resource optimization, into a single federated profile. This federated profile provides a comprehensive view of the system's resources and serves as a basis for making informed decisions on resource allocation and workload management. By leveraging the federated profile and employing FWP techniques, the experiment's objective was to evaluate the FWP in terms of memory, disk space, execution time, and other resource utilization parameters at the edge nodes like network statistics and the effects of ML algorithms on tree depth, and to demonstrate that it can aid with edge node selection and FPSN.

The following subsections describe the experimentation and the scenarios. It's worth noting that the same dataset and experimental setup used in Objective 3 were reused for this objective to maintain consistency and allow for direct comparisons. This approach ensures continuity and facilitates meaningful result discussions.

4.3.1 *Objective 4: Experimentation*

To evaluate the proposed model, four experimental scenarios were implemented.By conducting these experiments, we aimed to gain insights into the capabilities and limitations of the FWP model and to validate resource profiling optimization under different scenarios.

4.3.1.1 *Scenario 1. Federated Workload Profiling on Memory Usage*

Scenario 1 focuses on FWP's impact on memory usage, analyzing memory consumption before and after optimization across various ML models and edge nodes to identify the most memory-efficient ML model and node.

A. Best Performing ML Models with respect to Memory Usage

Initially, the effects of memory usage before and after FWP resource optimization techniques were examined using five different ML algorithms as shown in Figure 31.

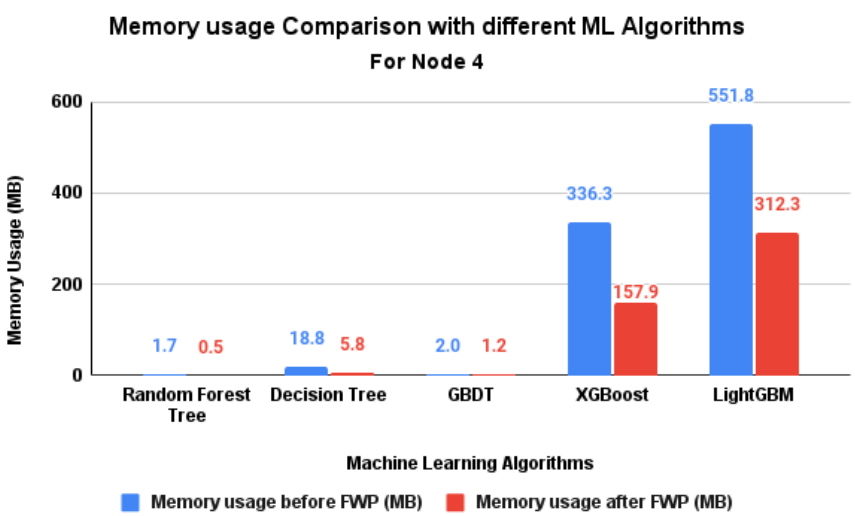


Figure 31: ML Algorithms Memory Usage Comparison

After implementing the FWP resource optimization algorithms, it was found through experimentation that memory consumption dropped significantly. Additionally, it can be deduced that Random Forest Tree, followed by GBDT and Decision Tree, were the best-performing ML models. The memory usage of LightGBM was the highest of all the models evaluated. In conclusion, the experiments revealed that Random Forest Tree was the top-performing ML model in terms of memory consumption, and thus the Random Forest model was employed for the rest of the experiments.

B. Memory usage prior to and after FWP at different edge nodes

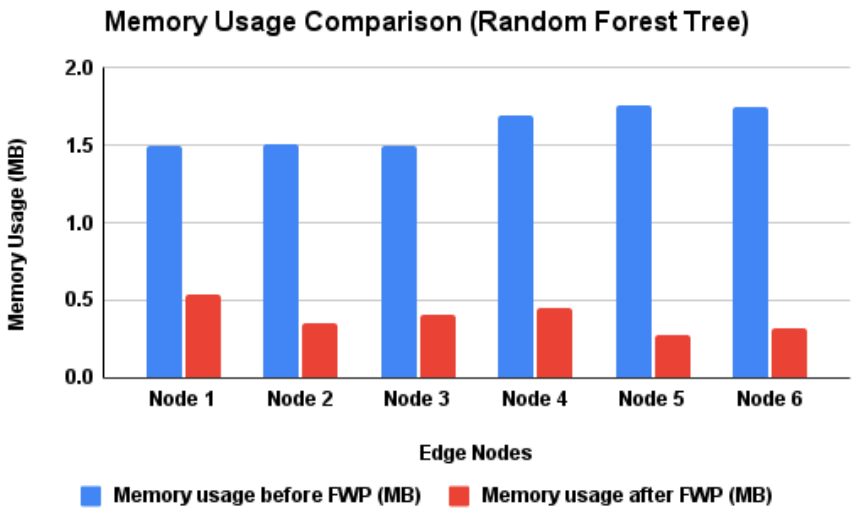


Figure 32: FWP Effect on Memory Usage on Nodes

FWP optimization strategies were implemented across six diverse nodes as seen in Figure 32, revealing a significant drop in the memory requirements in comparison with conventional approaches. Although all the

nodes benefited from FWP, Nodes 5, 6, and 2 saw the greatest reductions in the amount of memory they used respectively.

4.3.1.2 *Scenario 2. Federated Workload Profiling on Various Resource Parameters at The Node Level*

Scenario 2 explores FWP's impact on various resource parameters at the node level, specifically focusing on disk space usage and network I/O statistics before and after FWP optimization across different edge nodes.

A. Disk Space Usage Prior to and After FWP at Different Edge Nodes

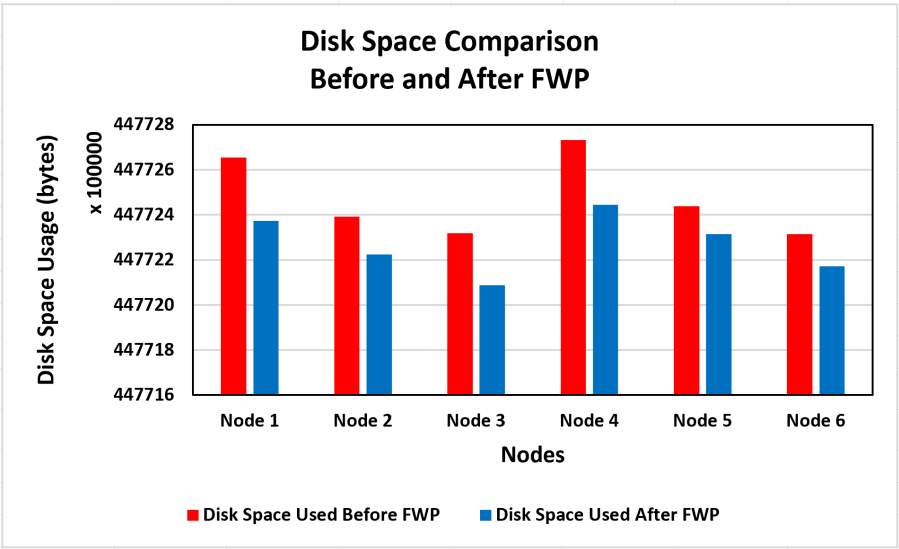


Figure 33: Impact of FWP on Disk Space

The FWP experiments' impact on disk space consumption was comparable with that of other previous experiments as evident from Figure 33. All of the nodes showed lower disk space usage after the FWP experiments, even though there was no large disparity in the usage of disk space.

B. Network I/O Statistics Prior to and After FWP

Each edge node's network I/O data were recorded, which included the characteristics bytes sent, bytes received, number of packets transmitted, number of packets received, total number of incoming dropped packets, and the total number of outbound lost packets.

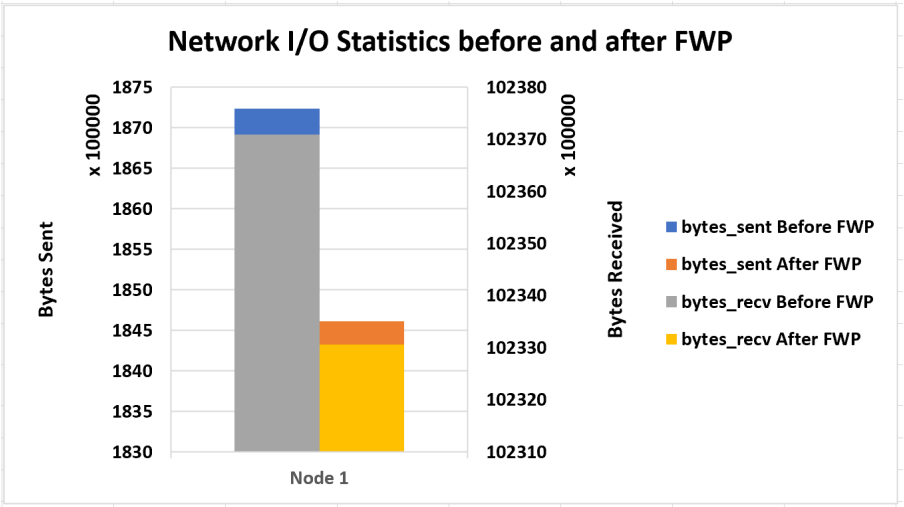


Figure 34: Effect of FWP on Network I/O

This I/O data analysis revealed crucial insights into the system's overall performance and helped identify potential bottlenecks and issues. After applying our proposed optimization strategies, which included resource-aware data reduction methods such as compression, protocol optimization, and caching, we were able to visually inspect the bytes sent and received and confirmed a decrease after implementing FWP, as shown in Figure 34.

4.3.1.3 Scenario 3. The Cumulative Impact of Federated Workload Profiling Across Expanding Node Networks

Extending the Scenario 2 experiments, an assessment was conducted to evaluate the cumulative impacts of optimization strategies on disk space, memory, and network performance as the number of nodes in the federation increases. The size of a federation can vary widely, ranging from a few nodes to hundreds or even thousands, depending on the specific network or system architecture. The number of nodes in a federation often impacts various aspects of network management, performance, and resource allocation.

A. FWP Impact on Cumulative Memory with Increasing Nodes

In this experiment, we explore the cumulative memory usage across ten nodes before and after implementing FWP and its impact as the number of nodes increases.

Figure 35 illustrates the cumulative memory usage across ten nodes both before and after the implementation of FWP. The blue line represents the memory consumption before FWP, and it shows a steady increase as additional nodes are incorporated into the system. Conversely, the orange line, depicting memory usage after FWP, displays a similar upward trend but with consistently lower values compared to the pre-FWP scenario. This suggests that FWP effectively optimizes memory resources, resulting in reduced memory utilization as the number of nodes in the federation increases. The graph highlights the positive impact of FWP on memory management, showcasing its ability to efficiently scale memory consumption with the growing number of nodes, potentially leading to enhanced system performance and resource allocation.

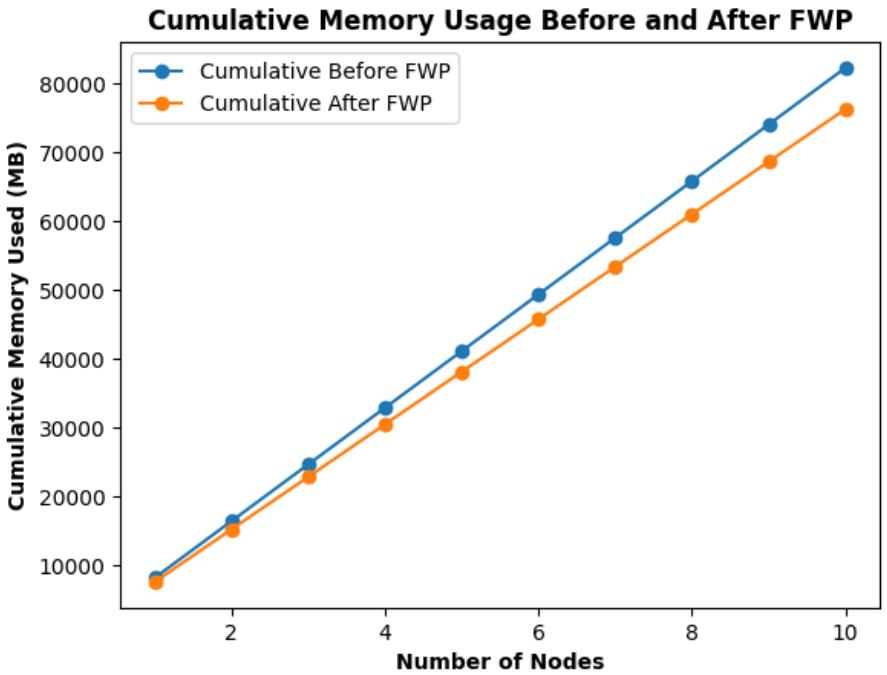


Figure 35: Cumulative Memory Usage Before and After FWP

B. FWP Effect on Cumulative Disk Space with Node Expansion

This experiment investigates the effects of FWP on cumulative disk space usage across an expanding node federation.

Figure 36 provides a visual representation of cumulative disk space usage across ten nodes, both before and after the implementation of FWP. The cumulative disk space usage is presented in gigabytes (GB) and is measured as the sum of disk space occupied by each node in the federation. The blue line represents the cumulative disk space consumption before FWP, showcasing the gradual increase in storage usage as more nodes are incorporated into the network. Each point on the line corresponds to a specific node, illustrating the progressive accumulation of data storage requirements across the network. In

contrast, the orange line represents cumulative disk space usage after FWP activation. There is a similar growth trend as more nodes join the network. The graph utilizes a dual-axis approach, with separate y-axes on the left and right sides, to present cumulative disk space usage before and after the implementation of FWP. This separation prevents data from overlapping with the other, ensuring that both trends can be easily examined and compared.

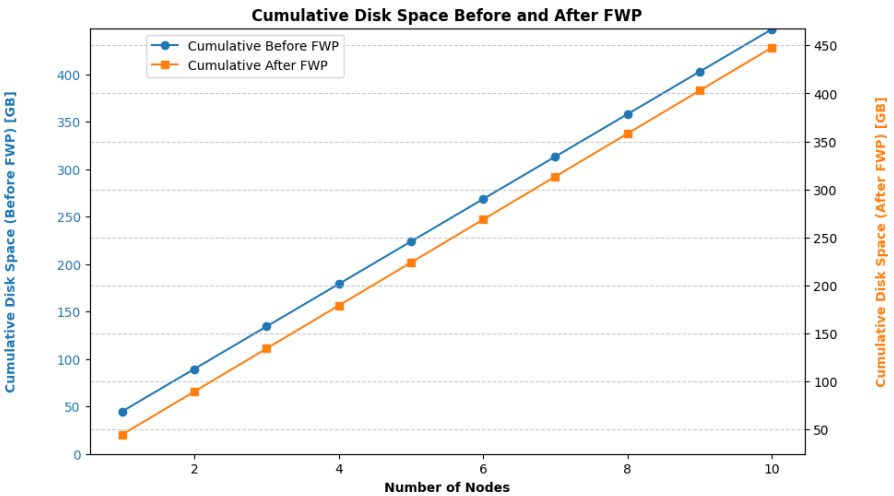


Figure 36: Cumulative Disk Space Consumption Before and After FWP

The graph suggests that the difference in cumulative disk space between before and after the implementation of FWP is generally negligible. In practical terms, this indicates that FWP has not significantly impacted the cumulative disk space usage across the ten nodes. Overall, the negligible difference in cumulative disk space suggests that FWP has moderately reduced but not substantially increased or decreased the total storage requirements of the system, at least within the observed scope of these ten nodes.

C. FWP's Influence on Network I/O with Increasing Federation Nodes

In this experiment, we investigate the effects of FWP on network I/O statistics as the number of nodes within the federation expands.

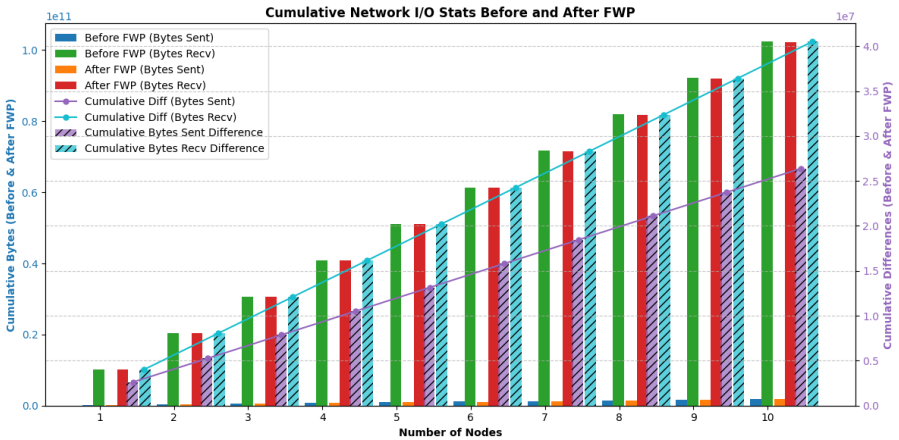


Figure 37: Cumulative Network I/O Stats Before and After FWP

Figure 37 offers a comprehensive perspective on network I/O (Input/Output) statistics for ten nodes, featuring four key components. The blue bars denote cumulative bytes sent before the introduction of FWP, while the green bars represent cumulative bytes received before FWP. Similarly, the orange bars depict cumulative bytes sent after FWP, and the red bars showcase cumulative bytes received after FWP. Additionally, the graph incorporates purple bars, which represent the cumulative difference in bytes sent, and cyan bars that represent the cumulative difference in bytes received. The two y-axes in the graph serve distinct purposes: the primary y-axis represents cumulative network traffic volumes, while the secondary y-axis illustrates the cumulative differences in network activity before and after the implementation of

Federated Workload Profiling (FWP). The cumulative difference bars on the secondary y-axis appear relatively higher, mainly because they are scaled using a factor to emphasize even subtle variations in network behavior. This comprehensive visualization aids in understanding the impact of FWP on network data flow across multiple nodes.

The graph demonstrates a notable network traffic pattern. Initially, there is a reduction in data bytes sent and received at Node 1 after implementing FWP. However, as more nodes join the federation, there is a consistent increase in network activity. Notably, the cumulative difference bars (purple and cyan) remain positive, highlighting FWP's ability to effectively manage and optimize data traffic as the network scales. This trend suggests significant network traffic savings when FWP is employed with a larger number of nodes, enhancing network efficiency and resource utilization.

4.3.1.4 Scenario 4. Federated Resource Profiling on Tree Depth

Scenario 3, investigates the impact of FWP on the convergence and memory utilization of machine learning algorithms.

A. Faster Convergence of ML Algorithms

When compared with the conventional approaches, the execution time of each of the five different machine learning algorithms evaluated with the proposed FWP was found to be significantly reduced, as shown in Figure 38. According to the results, Decision Tree and Random Forest performed better than other ML algorithms and converged more quickly. Even though the worst-performing model with the longest execution time was GBDT, our proposed model, FWP, cut the execution time to almost half.

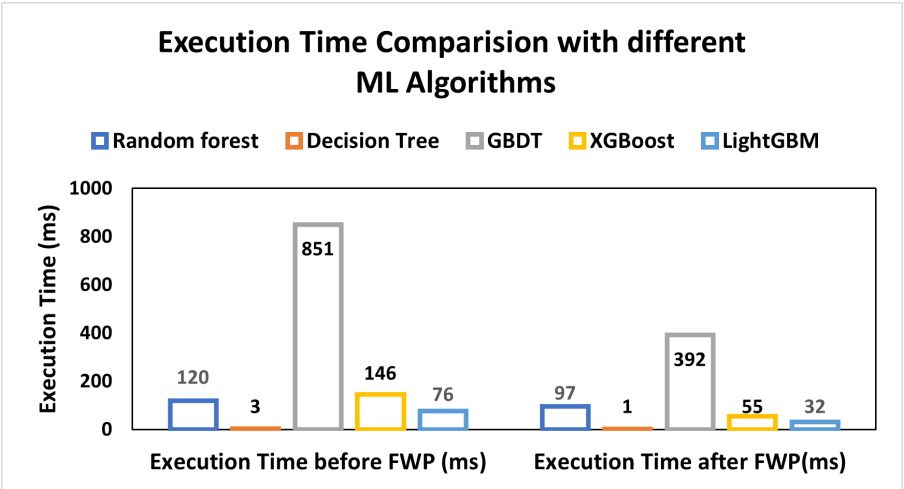


Figure 38: Execution Time Before and After FWP

B. Memory utilization

Algorithms for machine learning, such as Random Forest, have memory requirements that vary with the number and size of the trees in the algorithm. When dealing with big or complex datasets, the tree might be quite deep and include thousands of nodes. In such circumstances, the memory usage will increase exponentially. A decrease in the depth of the tree will lead to a reduction in the amount of memory that is used.

In Figure 39, the graph provides us with a visual representation of how the Random Forest Tree Depth affected execution time when FWP is utilized. It is evident that this was the reason for the decreased utilization of resources following FDP implementation.

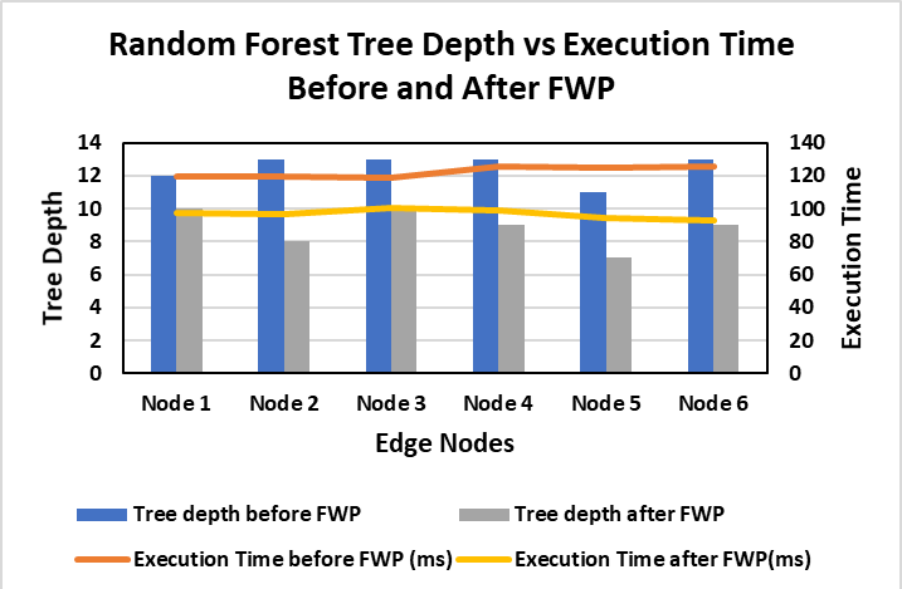


Figure 39: Effects of FWP on ML Tree Depth and Execution Time

4.3.1.5 *Scenario 5: Federated Workload Profiling aiding in Edge Node Selection and FPSN*

Depending on the application, the edge nodes can be mobile or static in edge computing. Mobility reduces the efficiency with which edges can exploit their resources, which can result in the FWP deciding on which edge nodes to be removed. To keep resource utilization and performance at their peak, the algorithm also takes into account the reputation of each node, keeping only the most trustworthy edge nodes even if they are mobile. The Response time; metric scores related to CPU, memory, network, sensors, processes, and storage; reputation score; mobility; etc. all factor into the weighted average score calculated at the server and used as a selection criterion for the EWP nodes. If two nodes have the same weighted average score, the server prioritizes the node with the highest reputation score, followed by the node with the lowest

response time, the lowest mobility, and the best CPU and memory availability and selects the best n edges. Ultimately, the weight given to each of the factors depended on the specific requirements of the system and the trade-offs between aspects such as performance, scalability, and cost. It's important to continuously evaluate and refine the approach based on real-world performance to ensure optimal results.

Because of its poor response time, limited mobility, excessive congestion, and low reputation, Edge Node 6 is removed after FWP is applied. Since time is of the essence in FPSN calculation, only the most resource-rich nodes were selected. To reduce the response time and to get a precise performance measurement while maintaining high accuracy, only trustworthy nodes were selected to use in our FPSN calculation by factoring in their reputation. It is evident that the performance measure accuracy has slightly improved after resource optimizations following FWP, which was applied at the nodes and used in FPSN calculations as shown in Figure 40.

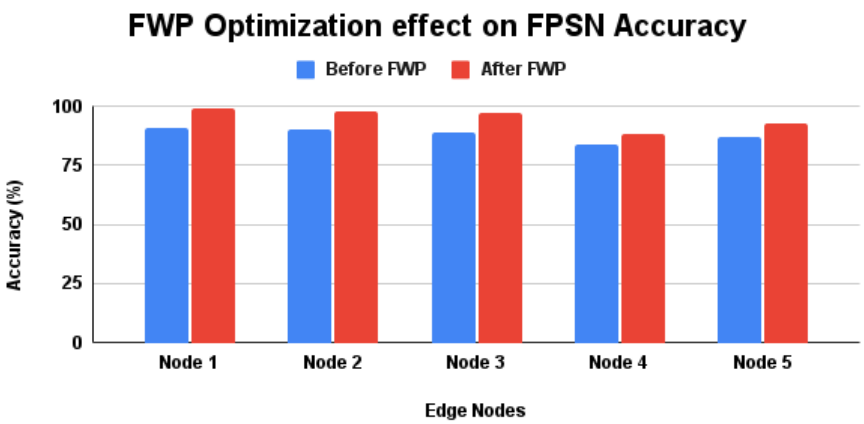


Figure 40: FWP Enabled FPSN Performance

4.3.2 Results Discussion

In this study, we presented four different scenarios that utilized FWP to optimize the resource utilization of edge computing nodes.

In Scenario 1, we evaluated the memory usage of five different machine learning models before and after applying FWP optimization techniques. On an average, applying FWP led to a 47.8% reduction in memory usage across all machine learning algorithms. The algorithms that benefited the most from FWP in terms of memory usage reduction were Random Forest Tree and Decision Tree, with reductions of 73.9% and 69.3%, respectively. XGBoost and LightGBM had the smallest reductions in memory usage, at 53% and 43.5%, respectively. We found that Random Forest Tree was the best-performing model in terms of memory consumption and used it for the rest of the experiments.

In Scenario 2, we analyzed the impact of FWP on disk space usage and network I/O statistics. Our proposed optimization strategies resulted in slightly lower disk space usage and decreased network I/O after implementing FWP. Specifically, the total disk space used decreased by 336,304 bytes. With the network I/O statistics analysis, we can see that the on an average the number of bytes sent by a node decreased by 2,623,071 bytes (or about 1.4%), while the number of bytes received by a node decreased by 40,266 bytes (or about 0.04%).

In Scenario 3, we focused on evaluating the impact of FWP on the convergence time and memory utilization of machine learning algorithms. Our results showed that the execution time of the evaluated algorithms was

significantly reduced, with Decision Tree and Random Forest performing the best in terms of faster convergence. The execution time decreased after FWP was applied, with percentage decreases ranging from 19.2% to 62.3%. Specifically, the execution times for Random Forest, Decision Tree, GBDT, XGBoost, and LightGBM decreased by 19.2%, 33.3%, 53.9%, 62.3%, and 57.9%, respectively. For the Random Forest tree depth and training time before and after FWP, we found that the average tree depth decreased by 3.34, or about 27.5%, after FWP was applied which will in turn lead to a reduction in the amount of memory. Similarly, the average execution time decreased by 24.68 ms, or about 20.4%, after FWP was applied. These results suggest that FWP had a significant positive impact on both tree depth and training time for Random Forest across all nodes.

Finally, in Scenario 4, we utilized FWP to aid in edge node selection and FPSN accuracy calculation. We removed an underperforming edge node and only selected trustworthy nodes to use in our FPSN calculation. On an average, applying FWP led to a 6.8% increase in accuracy across all nodes. However, it's worth noting that the degree of improvement varied across different nodes, with Node 1 having the largest increase in accuracy (8%), and Node 4 having the smallest increase (4%).

Overall, the results of our Objective 4 experiments demonstrate the effectiveness of FWP in optimizing the resource utilization of edge computing nodes. By methodically tailoring the utilization and selection of resources,we aimed to enhance performance, potentially improving the Quality of Edge Computing Services and operational efficiency in edge scenarios. The proposed optimization strategies were able to significantly reduce memory and disk space

usage, improve network I/O performance, and reduce the convergence time of machine learning algorithms. Furthermore, the FWP approach can be used to aid in edge node selection and improve the accuracy of FPSN.

4.4 Experimentation Summary

In our comprehensive experimentation, we rigorously evaluated the three proposed models. Each model was subjected to a series of detailed experiments designed to test their effectiveness in various scenarios.

The PSN Fusion Model excelled in handling clinical text data and cardiovascular diseases, outperforming standard algorithms with its ability to manage complex datasets and integrate static and dynamic data, showcasing robustness in diverse healthcare settings.

The FPSN Data Quality Aware Model, applied to fetal health assessment using cardiotocograms, significantly improved both data quality and prediction accuracy. The model excelled in scenarios demanding high data fidelity, showcasing its critical value in healthcare applications where accuracy is paramount.

The Resource Aware PSN Model focused on optimizing resource utilization, particularly in computing environments. It achieved notable reductions in memory and disk space usage, alongside improvements in network performance.

Together, these models offer innovative approaches to healthcare and resource management, demonstrating their effectiveness and versatility in addressing current and future challenges in data representation, quality assessment, and resource optimization across various fields.

Chapter 5: Conclusion and Future Perspective

This thesis has explored the domains of Smart Connected Health (SCH), Patient Similarity Networks (PSN), Federated Data Quality Profiling (FDQP), and Federated Workload Profiling (FWP). The research conducted has yielded significant findings and contributions in these areas.

In the study of SCH, we provided a comprehensive analysis of existing SCH solutions, considering SCH-enabling technologies, application contexts, and future developments. Additionally, we proposed an architectural model that encompasses the technological aspects and key stakeholders involved in SCH implementations. This work offers valuable insights and guidance for researchers and healthcare organizations seeking to implement and deploy SCH solutions.

The development of the PSN Fusion model has been another crucial aspect of this research. By integrating data mining and DL techniques, we proposed a multidimensional PSN model that effectively captures both contextual and longitudinal data. This model addresses the challenge of data heterogeneity and high dimensionality, resulting in more accurate patient similarity measures and personalized healthcare recommendations. Our experiments demonstrated that the DL-based PSN Fusion model outperformed traditional classification algorithms, highlighting its potential for improving patient outcomes.

The exploration of FDQ profiling emphasized the importance of data quality throughout the health data value chain. By incorporating data quality

metrics and feature selection techniques, we developed a framework that enhances the reliability and accuracy of healthcare data. This framework has the potential to significantly improve decision-making processes and support evidence-based healthcare practices.

The investigation into FWP aimed at optimizing resource utilization in edge computing nodes. Through experimentation, we demonstrated the effectiveness of FWP techniques in reducing memory usage, improving disk space usage, enhancing network I/O performance, and reducing convergence time in machine learning algorithms. Such enhancements are expected to contribute to a higher Quality of Edge Computing Services, especially regarding resource allocation and service responsiveness. However, a thorough evaluation of the overall Quality of Edge Computing Services based on these optimizations remains a subject for future research.

5.1 Contributions to the Field

This research has made several significant contributions to the field of healthcare and technology:

SCH: Our classification of existing SCH solutions and proposed architectural model provide researchers and healthcare organizations with a comprehensive framework for implementing and deploying SCH solutions. This work promotes the efficient transformation of the healthcare sector into an interconnected and data-driven ecosystem.

PSN Fusion Model: The development of the multidimensional PSN Fusion model offers a novel approach to patient similarity analysis, personalized healthcare recommendations, and improved patient outcomes. By

combining data mining and DL techniques, this model leverages the power of healthcare data to provide richer clinical evidence and support accurate diagnoses and treatment decisions.

FDQ Profiling: Our framework for FDQ profiling addresses the critical issue of data quality in healthcare settings. By integrating data quality metrics and feature selection techniques, this framework enhances the reliability and trustworthiness of healthcare data, enabling more robust decision-making processes and improving overall healthcare outcomes.

FWP: The investigation into FWP and the optimization of resource utilization in edge computing nodes contribute to the efficiency and cost-effectiveness of healthcare systems. The proposed techniques for reducing memory usage, improving disk space utilization, enhancing network I/O performance, and reducing convergence time in machine learning algorithms offer practical solutions for resource optimization in healthcare environments.

5.2 Research Strengths

The overall research exhibits several notable strengths. Firstly, the proposed models, such as the PSN Fusion Model and the federated resource profiling model, address important challenges in the healthcare domain, including patient representation and resource utilization optimization in a collaborative setting. These models showcase the research's ability to tackle real-world problems and provide practical solutions.

Secondly, the research demonstrates a strong experimental approach, conducting rigorous evaluations and analysis across multiple scenarios. The

experiments cover a wide range of factors, such as different datasets, similarity measures, distance evaluation approaches, and optimization techniques, providing a comprehensive assessment of the proposed models' performance.

Thirdly, the research incorporates advanced techniques and methodologies, including NLP models, BERT, autoencoders, and SNF algorithms, similarity distance measurement approaches, Federated Feature Selection, and FWP. By leveraging these state-of-the-art approaches, the research showcases a deep understanding of relevant technologies and their applicability to healthcare data analysis.

Fourthly, the research demonstrates a holistic perspective by considering both static and dynamic features, thereby capturing the complexity and temporal aspects of healthcare data. This comprehensive approach enables a more nuanced understanding of patient profiles and resource utilization patterns, enhancing the accuracy and practicality of the proposed models.

Lastly, the research emphasizes the significance of data privacy and quality. The inclusion of federated approaches and the evaluation of data quality metrics showcase the research's commitment to addressing privacy concerns and ensuring the reliability and integrity of the analyzed data.

Overall, the research exhibits strengths in its problem-solving approach, experimental approach, utilization of advanced techniques, consideration of different data aspects, and emphasis on privacy and data quality, making it a valuable contribution to the field of healthcare data analysis and personalized medicine.

5.3 Limitations

While the research exhibits several strengths, it is important to acknowledge a couple of limitations. Firstly, the evaluation and experimentation primarily relied on specific datasets and scenarios, which may not fully capture the complexity and diversity of real-world healthcare systems. Therefore, the generalizability of the findings to different contexts and datasets should be approached with caution. Secondly, the proposed models and methodologies may require further validation and refinement through larger-scale studies and real-world implementations. This would help assess their performance, scalability, and applicability in different healthcare settings, considering factors such as data privacy, data quality, and computational resources.

5.4 Real-World Challenges: Insights and Mitigation Strategies

As we combine the cutting-edge methodologies in SCH and PSN with the practical world of healthcare, we recognize the inherent challenges that come with bridging theoretical models and real-world implementations. These challenges, while daunting, offer us an opportunity to refine our strategies, ensuring they are not just theoretically sound but also pragmatically viable. Following are some key challenges to consider:

1. Resource Constraints in Edge Computing: Deploying models in edge computing environments requires a fine balance. While we aim for maximal accuracy and real-time processing, the hardware constraints of edge devices can be limiting. Memory restrictions, computational power, and network latency issues often necessitate compromises. For

instance, while a complex neural network might give superior results, its deployment might be untenable on a resource-limited device. Thus, achieving a balance between model complexity and feasibility of deployment remains a pivotal challenge.

Mitigation Strategy: Introduce adaptive algorithms that adjust computational tasks based on available resources. For edge nodes with significant constraints, lighter-weight models or algorithms can be employed to ensure seamless operation.

2. Data Privacy and Security Concerns: The federated nature of our methodologies inherently addresses some of the concerns related to data privacy. However, the very act of transferring insights, even if not raw data, can pose security threats. Beyond just privacy, there's the risk of data tampering, which can greatly skew results and lead to incorrect conclusions or medical recommendations. Ensuring robust encryption and authentication methods while minimizing exposure is a challenge that continues to evolve with the threat landscape.

   Mitigation Strategy: To address data privacy and security in real-world implementations, we advocate for the adoption of differential privacy, ensuring individual data remains obfuscated during analysis, and homomorphic encryption, which permits computations on encrypted data without decryption. Integrating blockchain technology can further enhance data integrity by providing a tamper-proof record of transactions and ensuring traceability.

3. Integration with Existing Systems: Our models, while innovative, need to integrate seamlessly with existing healthcare IT systems. These

systems, often legacy in nature, have their own nuances and idiosyncrasies. Ensuring our models are not just compatible but can also enhance the functionality of these systems is a challenge. Factors like data formats, communication protocols, and system-specific requirements mandate a flexible and adaptable approach to model deployment.

Mitigation Strategy: Adopting a modular model design and utilizing middleware solutions facilitates seamless integration with existing healthcare IT systems. Continuous collaboration with system stakeholders further ensures alignment and enhancement of legacy infrastructures.

4. Ethical Implications of AI-centric SCH Frameworks: As AI becomes increasingly embedded within SCH systems, there are thoughtful ethical considerations to address. AI-centric frameworks can potentially lead to automated decision-making processes in patient care, necessitating robust checks to prevent unintended biases that can lead to misdiagnoses or inappropriate treatments. Moreover, there's an ever-present risk of the technology being used in ways not originally intended, especially when proprietary algorithms are opaque and not open to scrutiny.

   Mitigation Strategy: To navigate the ethical challenges, it's crucial to adopt a transparent and participatory approach in the development and deployment of AI models. Engaging ethicists, clinicians, and patient representatives, in the design process can offer diverse insights and early identification of ethical issues. Continuous monitoring and auditing of AI models post-deployment can help in addressing unintended biases.

Embracing explainable AI principles can foster trust among end-users and stakeholders, while firm governance and standards can guide the ethical integration of AI in SCH.

The aforementioned challenges underscore the need for balance between innovation and practicality. As we progress, our goal remains to address these challenges, constantly refining our models and methodologies to meet the evolving demands of real-world healthcare settings.

5.5 Future Directions

While this research has made significant strides in the areas of SCH, PSN Fusion, FDQ profiling, and FWP, there are several avenues for future exploration:

Advanced DL Optimization: Further research is needed to optimize the DL neural networks used within the PSN Fusion model. Exploring advanced DL optimization techniques, such as network architecture search, automated hyperparameter tuning, and knowledge distillation, can enhance the accuracy and efficiency of the model.

Richer Datasource: Additionally, incorporating additional sources of data, such as genomic information or wearable device data, could provide richer patient datasets and enhance the overall effectiveness of the PSN model in personalized healthcare applications.

Emerging Technologies: In addition, integrating emerging technologies such as transfer learning or generative models could be explored to leverage pre-trained models and enhance the representation learning capabilities of PSN.

Intelligent DQ Enhancement: Future studies should focus on developing advanced algorithms and techniques to handle noisy, incomplete, and heterogeneous healthcare data effectively.

Dynamic Resource Allocation: Incorporating dynamic resource allocation strategies, such as reinforcement learning or adaptive algorithms, could optimize resource utilization in real-time based on changing workload and system demands, leading to improved efficiency and cost-effectiveness in healthcare environments.

Quality of Edge Computing Services: Future directions include refining strategies to enhance the real-time adaptability and efficiency of edge node resources, ensuring improved service quality in evolving computational environments.

Looking ahead, further exploration and refinement of the PSN model hold the key to pushing accuracy and efficiency to its maximum potential. This ongoing pursuit of optimization will continue to be a focal point in our efforts to enhance the PSN model's predictive capabilities, thereby enhancing its potential impact on precision medicine and the broader landscape of smart connected health.

References

[1] J. Thomason, "Big tech, big data and the new world of digital health," *Global Health Journal*, vol. 5, no. 4, pp. 165–168, 2021. doi:10.1016/j.glohj.2021.11.003.

[2] A. McBride, R. Hall, and K. O'Neill, "How can an intelligent health ecosystem create a smarter health experience?" [Accessed: 2023-10-09]. [Online]. Available: https://www.ey.com/en_gl/health/how-can-an-intelligent-health-ecosystem-create-a-smarter-health-experience.

[3] "Global digital health market forecast 2025 | Statista," [Accessed: 2023-03-21]. [Online]. Available: https://www.statista.com/statistics/1092869/global-digital-health-market-size-forecast/.

[4] "Healthcare Predictive Analytics Market | Research Report [2029]," [Accessed: 2023-07-24]. [Online]. Available: https://www.fortunebusinessinsights.com/healthcare-predictive-analytics-market-107352.

[5] C. C. Zwack, M. Haghani, M. Hollings, L. Zhang, S. Gauci, R. Gallagher, and J. Redfern, "The evolution of digital health technologies in cardiovascular disease research," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–11, 2023. doi:10.1038/s41746-022-00734-2.

[6] Deloitte, "Infographic: Global health care outlook 2021," *Deloitte*, 2021.

[7] K. Kumar, N. Kumar, and R. Shah, "Role of IoT to avoid spreading of COVID-19," *International Journal of Intelligent Networks*, 2020. doi:10.1016/j.ijin.2020.05.002.

[8] C. Morrison, S. Rimpiläinen, I. Bosnic, J. Thomas, and J. Savage, "Emerging Trends in Digital Health and Care: A Refresh Post-COVID," pp. 1–4, 2022. doi:10.17868/strath.00082203.

[9] J. Li and P. Carayon, "Health Care 4.0: A vision for smart and connected health care," *IISE Transactions on Healthcare Systems Engineering*, vol. 11, no. 3, pp. 171–180, 2021. doi:10.1080/24725579.2021.1884627.

[10] L. Rooney, S. Rimpiläinen, C. Morrison, and S. L. Nielsen, "Review of Emerging Trends in Digital Health and Care : A report by the Digital Health and Care Institute," no. November, 2019. doi:10.17868/67860.

[11] Soumitra Dutta, B. Lanvin, L. R. León, and S. Wunsch-Vincent, *Global Innovation Index 2022- What is the future of innovation- driven growth?*, 2022, vol. 2, no. 8.5.2017. ISBN 9789280534320.

[12] K. P. Fadahunsi, S. O'Connor, J. T. Akinlua, P. A. Wark, J. Gallagher, C. Carroll, J. Car, A. Majeed, and J. O'Donoghue, "Information quality frameworks for digital health technologies: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 5, pp. 1–12, 2021. doi:10.2196/23479.

[13] C. M. Clancy, "Getting To 'Smart' Health Care," *Health Affairs*, vol. 25, no. Suppl1, 1 2006. doi:10.1377/hlthaff.25.w589.

[14] "National Science Foundation, "Smart and Connected Health (SCH) Program Solicitation NSF 18-541," p. 1–17, 2013.

[15] K. Taylor, "Connected health - how digital technology is transforming health and social care," *Deloitte*, p. 40, 2015.

[16] M. Chen, J. Qu, Y. Xu, and J. Chen, "Smart and connected health : What can we learn from funded projects ?" *Data Inf. Manag*, vol. 1, no. 2, 2018.

[17] R. Vaishya, M. Javaid, I. Khan, and A. Haleem, "Artificial intelligence (ai) applications for covid-19 pandemic," *Diabetes Metab. Syndr. Clin. Res. Rev*, vol. 14, no. 4, pp. 337–339,, 2020.

[18] M. Z. Nezhad, D. Zhu, N. Sadati, K. Yang, and P. Levi, "Subic: A supervised bi-clustering approach for precision medicine," in *2017 16th IEEE International Conference on Machine Learning and Applications ICMLA*. IEEE, 2017, pp. 755–760.

[19] S. F. Terry, "Obama's Precision Medicine Initiative," *Genetic Testing and Molecular Biomarkers*, vol. 19, no. 3, pp. 113–114, 2015. doi:10.1089/gtmb.2015.1563.

[20] "Emirati Genome Progam," [Accessed: 2023-08-09]. [Online]. Available: https://emiratigenomeprogram.ae/.

[21] F. Du, C. Plaisant, N. Spring, and B. Shneiderman, "Finding Similar People to Guide Life Choices," pp. 5498–5544, 2017. doi:10.1145/3025453.3025777.

[22] "PatientsLikeMe," [Accessed: 2023-04-22]. [Online]. Available: https://www.patientslikeme.com/.

[23] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England journal of medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[24] S.-A. Brown, "Patient Similarity: Emerging Concepts in Systems and Precision Medicine," *Frontiers in Physiology*, vol. 7, no. November, pp. 1–6, 2016. doi:10.3389/fphys.2016.00561.

[25] H. S. A. Fang, N. C. Tan, W. Y. Tan, R. W. Oei, M. L. Lee, and W. Hsu, "Patient similarity analytics for explainable clinical risk prediction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–12, 2021. doi:10.1186/s12911-021-01566-y.

[26] S. Pai and G. D. Bader, "Patient Similarity Networks for Precision Medicine," *Journal of Molecular Biology*, vol. 430, no. 18, pp. 2924–2938, 2018. doi:10.1016/j.jmb.2018.05.037.

[27] E. Parimbelli, S. Marini, L. Sacchi, and R. Bellazzi, "Patient similarity for precision medicine: A systematic review," *Journal of Biomedical Informatics*, vol. 83, no. January, pp. 87–96, 2018. doi:10.1016/j.jbi.2018.06.001.

[28] I. Tobore, J. Li, L. Yuhang, Y. Al-Handarish, A. Kandwal, Z. Nie, and L. Wang, "Deep learning intervention for health care challenges: Some biomedical domain considerations," *Journal of Medical Internet Research*, vol. 21, no. 8, 2019. doi:10.2196/11966.

[29] A. N. Navaz, E. Mohammed, M. A. Serhani, and N. Zaki, "The use of data mining techniques to predict mortality and length of stay in an ICU," in *2016 12th International Conference on Innovations in Information Technology (IIT)*. IEEE, 11 2016. doi:10.1109/INNOVATIONS.2016.7880045. ISBN 978-1-5090-5341-4.

[30] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, 12 2019. doi:10.1038/s41597-019-0103-9.

[31] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, 12 2019. doi:10.2200/S00960ED2V01Y201910AIM043.

[32] R. Fantacci and B. Picano, "Federated learning framework for mobile edge computing networks," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 15–21, 2020. doi:10.1049/trit.2019.0049.

[33] B. Pfitzner, N. Steckhan, and B. Arnrich, "Federated Learning in a Medical Context : A Systematic Literature Review," vol. 21, no. 2, 2021. doi:10.1145/3412357.

[34] P. M. Mammen, "Federated Learning: Opportunities and Challenges," 2021. doi:10.48550/arXiv.2101.05428.

[35] M. Mirzaie, B. Behkamal, and S. Paydar, "Big Data Quality: A systematic literature review and future research directions," *arXiv*, 2019. doi:10.48550/arXiv.1904.05353.

[36] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, *Big data quality framework: a holistic approach to continuous quality management*. Springer International Publishing, 2021, vol. 8, no. 1. ISBN 4053702100468.

[37] M. Ashouri, P. Davidsson, and R. Spalazzese, "Quality attributes in edge computing for the Internet of Things: A systematic mapping study," *Internet of Things (Netherlands)*, vol. 13, p. 100346, 2021. doi:10.1016/j.iot.2020.100346.

[38] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, 2023. doi:10.1109/JPROC.2022.3226481.

[39] F. Zhang, H. Yan, and X. Zhang, "QoE-Aware User Allocation in Edge Computing Systems for Popular Services," *Proceedings - 2022 8th International Conference on Big Data and Information Analytics, BigDIA 2022*, no. BigDIA, pp. 226–231, 2022. doi:10.1109/BigDIA56350.2022.9874194.

[40] M. Chowdhury, "Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques," *Sensors (Switzerland)*, vol. 20, no. 11, 2020. doi:10.3390/s20113127.

[41] A. Haq, "Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data." doi:10.3390/s20092649.

[42] K. Chung, H. Yoo, and D.-E. Choe, "Ambient context-based modeling for health risk assessment using deep neural network," *J. Ambient Intell. Humaniz. Comput*, vol. 11, no. 4, pp. 1387–1395,, 2020. doi:10.1007/s12652-018-1033-7.

[43] R. Aileni, S. Pasca, and A. Florescu, "Eeg-brain activity monitoring and predictive analysis of signals using artificial neural networks," *Sensors*, 2020. doi:10.3390/s20123346.

[44] M. Chand, N. Ramachandran, D. Stoyanov, and L. Lovat, "Robotics, artificial intelligence and distributed ledgers in surgery: data is key!" *Tech. Coloproctol*, vol. 22, no. 9, pp. 645–648,, 2018. doi:10.1007/s10151-018-1847-5.

[45] S. Mohammed, H. Park, C. Park, Y. Amirat, and B. Argall, "Special issue on assistive and rehabilitation robotics," *Autonomous Robots*, vol. 41, no. 3, pp. 513–517. doi:10.1007/s10514-017-9627-z.

[46] G. Ciuti, R. Caliò, D. Camboni, L. Neri, F. Bianchi, A. Arezzo, A. Koulaouzidis, S. Schostek, D. Stoyanov, C. M. Oddo, B. Magnani, A. Menciassi, M. Morino, M. O. Schurr, and P. Dario, "Frontiers of robotic endoscopic capsules: a review," *Journal of Micro-Bio Robotics*, vol. 11, no. 1-4, pp. 1–18. doi:10.1007/s12213-016-0087-x.

[47] A. E. Chung, A. C. Griffin, D. Selezneva, and D. Gotz, "Health and fitness apps for hands-free voice-activated assistants: Content analysis," *JMIR mHealth and uHealth*, vol. 6, no. 9, pp. 1–13, 2018. doi:10.2196/mhealth.9705.

[48] C. De Cock, M. Milne-Ives, M. Van Velthoven, A. Alturkistani, C. Lam, and E. Meinert, "Effectiveness of conversational agents (virtual assistants) in health care: Protocol for a systematic review," *JMIR Research Protocols*, vol. 9, no. 3, 2020. doi:10.2196/16934.

[49] S. Ginnavaram, M. Myneni, and B. Padmaja, "An intelligent assistive vr tool for elderly people with mild cognitive impairment: Vr components and applications," *Int. J. Adv. Sci. Technol*, vol. 29, no. 4, pp. 796–803,, 2020.

[50] Y. Kim, H. Kim, and Y. O. Kim, "Virtual reality and augmented reality in plastic surgery: A review," *Archives of Plastic Surgery*, vol. 44, no. 3, pp. 179–187, 2017. doi:10.5999/aps.2017.44.3.179.

[51] P. Philip, J. A. Micoulaud-Franchi, P. Sagaspe, E. D. Sevin, J. Olive, S. Bioulac, and A. Sauteraud, "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders," *Scientific Reports*, vol. 7, no. February, pp. 1–7, 2017. doi:10.1038/srep42656.

[52] P. Sanz Leon, S. A. Knock, M. M. Woodman, L. Domide, J. Mersmann, A. R. McIntosh, and V. Jirsa, "The virtual brain: a simulator of primate brain network dynamics," *Frontiers in Neuroinformatics*, vol. 7, 2013. doi:10.3389/fninf.2013.00010.

[53] H. Aerts, M. Schirner, T. Dhollander, B. Jeurissen, E. Achten, D. Van Roost, P. Ritter, and D. Marinazzo, "Modeling brain dynamics after tumor resection using the virtual brain," *NeuroImage*, vol. 213, no. March, p. 116738, 2020. doi:10.1016/j.neuroimage.2020.116738.

[54] A. Alexander, M. McGill, A. Tarasova, C. Ferreira, and D. Zurkiya, "Scanning the future of medical imaging," *Journal of the American College of Radiology*, vol. 16, no. 4, pp. 501–507, 2019. doi:10.1016/j.jacr.2018.09.050.

[55] N. Deepa, P. B, P. K. Reddy, R. G. Thippa, B. Thar, A. Khan, and U. Tariq, "An ai-based intelligent system for healthcare analysis using ridge–adaline stochastic gradient descent classifier," *Journal of Supercomputing*. doi:10.1007/s11227-020-03347-2.

[56] M. Abo-Tabik, N. Costen, J. Darby, and Y. Benn, "Towards a smart smoking cessation app: A 1d-cnn model predicting smoking events," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–18, 2020. doi:10.3390/s20041099.

[57] L. Romeo, A. Petitti, and R. Marani, "Internet of robotic things in smart domains : Applications and challenges," *Sensors*, 2020. doi:10.3390/s20123355.

[58] A. O. Andrade, A. A. Pereira, S. Walter, R. Almeida, R. Loureiro, D. Compagna, and P. J. Kyberd, "Bridging the gap between robotic technology and health care," *Biomedical Signal Processing and Control*, vol. 10, no. 1, pp. 65–78, 2014. doi:10.1016/j.bspc.2013.12.009.

[59] M. Antico, F. Sasazawa, L. Wu, A. Jaiprakash, J. Roberts, R. Crawford, A. K. Pandey, and D. Fontanarosa, "Ultrasound guidance in minimally invasive robotic procedures," *Medical Image Analysis*, vol. 54, pp. 149–167, 2019. doi:10.1016/j.media.2019.01.002.

[60] N. Ouerhani, A. Maalel, and H. Ben Ghézela, "Spececa: a smart pervasive chatbot for emergency case assistance based on cloud computing," *Cluster Computing*, vol. 23, p. 2471–2482, 2019. doi:10.1007/s10586-019-03020-1.

[61] B. W. Munzer, M. M. Khan, B. Shipman, and P. Mahajan, "Augmented reality in emergency medicine: A scoping review," *Journal of Medical Internet Research*, vol. 21, no. 4, pp. 1–10, 2019. doi:10.2196/12368.

[62] S. Wang, M. Parsons, J. Stone-McLean, P. Rogers, S. Boyd, K. Hoover, O. Meruvia-Pastor, M. Gong, and A. Smith, "Augmented reality as a telemedicine platform for remote procedural training," *Sensors (Switzerland)*, vol. 17, no. 10, pp. 1–21, 2017. doi:10.3390/s17102294.

[63] P. Macedo, C. Pereira, P. Mota, D. Silva, and A. Frade, "Conversational agent in mhealth to empower people managing the parkinson ' s disease," *Procedia Comput. Sci*, vol. 160, pp. 402–408,, 2019. doi:10.1016/j.procs.2019.11.074.

[64] C. A. J. Luis, Z. Montenegro and R. Righi, "Survey of conversational agents in health," *Expert Syst. Appl*, vol. 129, pp. 56–67,, 2019. doi:10.1016/j.eswa.2019.03.054.

[65] A. Davoudi, "Intelligent icu for autonomous patient monitoring using pervasive sensing and deep learning," *Sci. Rep*, 2019. doi:10.1038/s41598-019-44004-w.

[66] T. Steenkiste, "Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks," *Artif. Intell. Med*, 2019. doi:10.1016/j.artmed.2018.10.008.

[67] M. Bonaccorsi, L. Fiorini, F. Cavallo, A. Saffiotti, and P. Dario, "A cloud robotics solution to improve social assistive robots for active and healthy aging," *Int. J. Soc. Robot*, vol. 8, no. 3, pp. 393–408,, 2016. doi:10.1007/s12369-016-0351-1.

[68] B. Clipper, J. Batcheller, A. Thomaz, and A. Rozga, "Artificial intelligence and robotics: A nurse leader's primer," *Nurse Lead*, vol. 16, no. 6, pp. 379–384,, 2018. doi:10.1016/j.mnl.2018.07.015.

[69] I. Deutsch, H. Erel, M. Paz, G. Hoffman, and O. Zuckerman, "Home robotic devices for older adults: Opportunities and concerns," *Comput. Human Behav*, 2019. doi:10.1016/j.chb.2019.04.002.

[70] N. Hata, P. Moreira, and G. Fischer, "Robotics in mri-guided interventions," in *Topics in Magnetic Resonance Imaging*, 2018. doi:10.1097/RMR.0000000000000159.

[71] Y. Kassahun, "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *Int. J. Comput. Assist. Radiol. Surg*, vol. 11, no. 4, pp. 553–568,, 2016. doi:10.1007/s11548-015-1305-z.

[72] A. Khan, F. Bibi, M. Dilshad, S. Ahmed, Z. Ullah, and H. Ali, "Accident detection and smart rescue system using android smartphone with real-time location tracking," *Int. J. Adv. Comput. Sci. Appl*, vol. 9, no. 6, pp. 341–355,, 2018. doi:10.14569/IJACSA.2018.090648.

[73] S. Khan and A. Tsung, "Aso author reflections: The evolution of minimally invasive liver surgery and the future with robotics," *Ann. Surg. Oncol*, vol. 25, no. S3, pp. 786–787,, 2018. doi:10.1245/s10434-018-6880-0.

[74] D. Kumar, K. Achuthan, B. Nair, and S. Diwakar, "Online bio-robotics labs: Open hardware models and architecture," *Int. Conf. Robot. Autom. Humanit. Appl. RAHA*, pp. 1–5,, 2016. doi:10.1109/RAHA.2016.7931877.

[75] J. Lafuente, "Robotics assistance in neurosurgery—improving the outcome for our patients," *Acta Neurochir*, vol. 160, no. 10, pp. 1889–1889,, 2018. doi:10.1007/s00701-018-3624-7.

[76] F. Lanza, V. Seidita, and A. Chella, "Agents and robots for collaborating and supporting physicians in healthcare scenarios," *J. Biomed. Inform*, 2020. doi:10.1016/j.jbi.2020.103483.

[77] S. Mansouri, F. Farahmand, G. Vossoughi, and A. Ghavidel, "A comprehensive multimodality heart motion prediction algorithm for robotic-assisted beating heart surgery," *Int. J. Med. Robot. Comput. Assist. Surg*, vol. 15, no. 2, pp. 1–12,, 2019. doi:10.1002/rcs.1975.

[78] R. Miller and R. Miller, "Robotics and health care," *Ind. Robot Handb*, pp. 596–600,, 2013. doi:10.1007/978-1-4684-6608-9_68.

[79] B. Parsley, "Robotics in orthopedics: A brave new world," *J. Arthroplasty*, vol. 33, no. 8, pp. 2355–2357,, 2018. doi:10.1016/j.arth.2018.02.032.

[80] B. Teoh, "Utilising tatme and robotics to reduce r1 risk in locally advanced rectal cancer with rectovaginal and cervical involvement," *Tech. Coloproctol*, 2019. doi:10.1007/s10151-019-01941-y.

[81] Y. Weng and Y. Hirata, "Ethically aligned design for assistive robotics," *Conf. Intell. Saf. Robot. ISR*, pp. 286–290,, 2018. doi:10.1109/IISR.2018.8535889.

[82] D. Dimitrov, "Medical internet of things and big data in healthcare," *Healthc. Inform. Res*, vol. 22, no. 3, pp. 156–163,, 2016. doi:10.4258/hir.2016.22.3.156.

[83] P. Dineshkumar, R. Senthilkumar, K. Sujatha, R. Ponmagal, and V. Rajavarman, "Big data analytics of iot based health care monitoring system," *Int. Conf. Electr. Comput. Electron. Eng. UPCON*, pp. 55–60,, 2016. doi:10.1109/UPCON.2016.7894624.

[84] L. Greco, G. Percannella, P. Ritrovato, F. Tortorella, and M. Vento, "Trends in iot based solutions for health care: moving ai to the edge," *Pattern Recognit. Lett*, vol. 135, pp. 346–353,, 2020. doi:10.1016/j.patrec.2020.05.016.

[85] P. Moore and H. Pham, "A fog computing model for pervasive connected healthcare in smart environments," *Int. J. Grid Util. Comput*, vol. 10, no. 4, p. 375–391, 2019.

[86] A. Abdellatif, A. Mohamed, C. Chiasserini, M. Tlili, and A. Erbad, "Edge computing for smart health: Context-aware approaches, opportunities, and challenges," *IEEE Netw*, vol. 33, no. 3, pp. 196–203,, 2019. doi:10.1109/MNET.2019.1800083.

[87] D. Sangeetha, M. Rathnam, R. Vignesh, J. Chaitanya, and V. Vaidehi, "Medidrone—a predictive analytics-based smart healthcare system," in *Smart Innovation, Systems and Technologies*, vol. 164, 2020. doi:10.1007/978-981-32-9889-7_2 p. 19–33.

[88] M. Balasingam, "Drones in medicine—the rise of the machines," *Int. J. Clin. Pract*, 2017. doi:10.1111/ijcp.12989.

[89] K. Mandl, D. Gottlieb, and A. Ellis, "Beyond one-off integrations: A commercial, substitutable, reusable, standards-based, electronic health record-connected app," *J. Med. Internet Res*, vol. 21, no. 2, pp. 12 902,, 2019. doi:10.2196/12902.

[90] C. Maramis, “A smartphone application for semi-controlled collection of objective eating behavior data from multiple subjects,” *Comput. Methods Programs Biomed*, vol. 194, 2020. doi:10.1016/j.cmpb.2020.105485.

[91] I. Kamil and S. Ogundoyin, “A lightweight clas scheme with complete aggregation for healthcare mobile crowdsensing,” *Comput. Commun*, vol. 147, no. August, pp. 209–224,, 2019. doi:10.1016/j.comcom.2019.08.027.

[92] R. Kraft, “Efficient processing of geospatial mhealth data using a scalable crowdsensing platform,” *Sensors (Switzerland)*, vol. 20, no. 12, pp. 1–21,, 2020. doi:10.3390/s20123456.

[93] H. Jin, L. Su, H. Xiao, and K. Nahrstedt, “Incentive mechanism for privacy-aware data aggregation in mobile crowd sensing systems,” *IEEE/ACM Trans. Netw*, 2018. doi:10.1109/TNET.2018.2840098.

[94] H. Jin, L. Su, and K. Nahrstedt, “Centurion: Incentivizing multi-requester mobile crowd sensing,” in *Proceedings - IEEE INFOCOM*, 2017. doi:10.1109/INFOCOM.2017.8057111.

[95] M. Hassanalieragh, “Health monitoring and management using internet-of-things (iot) sensing with cloud-based processing: Opportunities and challenges,” in *Proc. - 2015 IEEE Int. Conf. Serv. Comput. SCC 2015*, 2015, pp. 285–292,.

[96] N. Mehta, A. Pandit, and M. Kulkarni, “Elements of healthcare big data analytics,” 2020. doi:10.1007/978-3-030-31672-3_2 pp. 23–43,.

[97] R. Iqbal, F. Doctor, and B. More, “Big data analytics: Computational intelligence techniques and application areas technological,” 2018. doi:10.1016/j.techfore.2018.03.024.

[98] R. Peckham and R. Sinha, “Anarchitectures of health: Futures for the biomedical drone,” *Glob. Public Health*, 2019. doi:10.1080/17441692.2018.1546335.

[99] K. Laksham, “Unmanned aerial vehicle (drones) in public health: A swot analysis,” *J. Fam. Med. Prim. Care*, 2019. doi:10.4103/jfmpc.jfmpc_413_18.

[100] F. Amato, S. Marrone, V. Moscato, G. Piantadosi, A. Picariello, and C. Sansone, “Holmes: ehealth in the big data and deep learning era,” *Inf*, vol. 10, no. 2, 2019. doi:10.3390/info10020034.

[101] M. Hossain and G. Muhammad, “Emotion-aware connected healthcare big data towards 5g,” *IEEE Internet Things J*, vol. 5, no. 4, pp. 2399–2406,, 2018. doi:10.1109/JIOT.2017.2772959.

[102] C. Cosgriff, L. Celi, and D. Stone, “Critical care, critical data,” *Biomed. Eng. Comput. Biol*, 2019. doi:10.1177/1179597219856564.

[103] G. Marques and R. Pitarma, “Occupational health and enhanced living environments through internet of things,” *RISTI - Rev. Iber. Sist. e Tecnol. Inf*, no. 19, pp. 1–13,, 2019.

[104] F. Firouzi, B. Farahani, M. Ibrahim, and K. Chakrabarty, “Keynote paper: From eda to iot ehealth: Promises, challenges, and solutions,” *IEEE Trans. Comput. Des. Integr. Circuits Syst*, vol. 37, no. 12, pp. 2965–2978,, 2018. doi:10.1109/TCAD.2018.2801227.

[105] L. Catarinucci, “An iot-aware architecture for smart healthcare systems,” *IEEE Internet Things J*, vol. 2, no. 6, pp. 515–526,, 2015. doi:10.1109/JIOT.2015.2417684.

[106] M. Aldeer, R. Martin, and R. Howard, “Pillsense: Designing a medication adherence monitoring system using pill bottle-mounted wireless sensors,” in *2018 IEEE Int. Conf. Commun. Work. ICC Work*, 2018. doi:10.1109/ICCW.2018.8403547 pp. – , 1–6,.

[107] M. Alaziz, Z. Jia, J. Liu, R. Howard, Y. Chen, and Y. Zhang, “Motion scale: A body motion monitoring system using bed-mounted wireless load cells,” *Proc. - 2016 IEEE 1st*, pp. 183–192,, 2016. doi:10.1109/CHASE.2016.13.

[108] I. Saied and S. Hussainy, “Portable and wearable device for microwave head diagnostic systems,” in *2019 IEEE Healthc. Innov. Point Care Technol. HI-POCT 2019*, 2019. doi:10.1109/HI-POCT45284.2019.8962890 pp. 45–48,.

[109] A. Alqadami, K. Bialkowski, A. Mobashsher, and A. Abbosh, "Wearable electromagnetic head imaging system using flexible wideband antenna array based on polymer technology for brain stroke diagnosis," *IEEE Trans. Biomed. Circuits Syst*, vol. 13, no. 1, pp. 124–134,, 2019. doi:10.1109/TBCAS.2018.2878057.

[110] R. Gupta, S. Member, S. Tanwar, S. Tyagi, and N. Kumar, "Habits : Blockchain-based telesurgery framework," *Conf. Comput. Inf. Telecommun. Syst*, pp. 1–5,, 2019. doi:10.1109/CITS.2019.8862127.

[111] A. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for iot," *Sensors (Switzerland)*, vol. 19, no. 2, pp. 1–17,, 2019. doi:10.3390/s19020326.

[112] A. Bhawiyuga, A. Wardhana, K. Amron, and A. Kirana, "Platform for integrating internet of things based smart healthcare system and blockchain network," in *Proc. - 2019 6th NAFOSTED Conf. Inf. Comput. Sci. NICS 2019*, 2019. doi:10.1109/NICS48868.2019.9023797 pp. 55–60,.

[113] S. Islam, D. Kwak, M. Kabir, M. Hossain, and K. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708,, 2015. doi:10.1109/ACCESS.2015.2437951.

[114] J. Long and J. Ehrenfeld, "The role of augmented intelligence (ai) in detecting and preventing the spread of novel coronavirus," *J. Med. Syst*, vol. 44, no. 3, pp. 2019–2020,, 2020. doi:10.1007/s10916-020-1536-6.

[115] V. G.-D. F. López-Martínez, E. R. Núñez-Valdez and Z. Bursac, "A case study for a big data and machine learning platform to improve medical decision support in population health management," *Algorithms*, vol. 13, no. 4, pp. 1–19,, 2020. doi:10.3390/A13040102.

[116] H. Wang, M. Daneshmand, and H. Fang, "Artificial intelligence (ai) driven wireless body area networks: Challenges and directions," *Proc. - IEEE Int. Conf. Ind*, no. Icii, pp. 428–429,, 2019. doi:10.1109/ICII.2019.00079.

[117] F. Fotouhi-Ghazvini and S. Abbaspour, "Wearable wireless sensors for measuring calorie consumption," *J. Med. Signals Sens*, vol. 10, no. 1, pp. 19–34,, 2020. doi:10.4103/jmss.JMSS_15_18.

[118] M. Boulos, J. Wilson, and K. Clauson, "Geospatial blockchain: Promises, challenges, and scenarios in health and healthcare," *Int. J. Health Geogr*, vol. 17, no. 1, pp. 1–10,, 2018. doi:10.1186/s12942-018-0144-x.

[119] T. Dey, S. Jaiswal, S. Sunderkrishnan, and N. Katre, "Healthsense: A medical use case of internet of things and blockchain," *Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017*, no. Iciss, pp. 486–491,, 2018. doi:10.1109/ISS1.2017.8389459.

[120] R. M. Aileni, S. Pasca, and A. Florescu, "EEG-Brain Activity Monitoring and Predictive Analysis of Signals Using Artificial Neural Networks," *Sensors*, 2020. doi:10.3390/s20123346.

[121] A. Bagula, M. Mandava, and H. Bagula, "A framework for healthcare support in the rural and low income areas of the developing world," *J. Netw. Comput. Appl*, vol. 120, pp. 17–29,, 2018. doi:10.1016/j.jnca.2018.06.010.

[122] J. Ramkumar, M. Baskar, P. Nipun, and A. Aithagani, "Effective framework to monitor patient health care through intelligent system," *Int. J. Adv. Sci. Technol*, vol. 29, no. 4, pp. 1828–1835,, 2020.

[123] N. Garg, M. Wazid, A. Das, D. Singh, J. Rodrigues, and Y. Park, "Bakmp-iomt: Design of blockchain enabled authenticated key management protocol for internet of medical things deployment," *IEEE Access*, vol. 8, pp. 95 956–95 977,, 2020. doi:10.1109/ACCESS.2020.2995917.

[124] R. Wang, H. Liu, H. Wang, Q. Yang, and D. Wu, "Distributed security architecture based on blockchain for connected health: Architecture, challenges, and approaches," *IEEE Wirel. Commun*, vol. 26, no. 6, pp. 30–36,, 2019. doi:10.1109/MWC.001.1900108.

[125] F. Bublitz, "Disruptive technologies for environment and health research: An overview of artificial intelligence, blockchain, and internet of things," *Int. J. Environ. Res. Public Health*, vol. 16, no. 20, pp. 1–24,, 2019. doi:10.3390/ijerph16203847.

[126] S. Amin, M. Hossain, G. Muhammad, M. Alhussein, and M. Rahman, "Cognitive smart healthcare for pathology detection and monitoring," *IEEE Access*, vol. 7, pp. 10 745–10 753,, 2019. doi:10.1109/ACCESS.2019.2891390.

[127] K. Azghiou, M. El Mouhib, M. A. Koulali, and A. Benali, "An End-to-End Reliability Framework of the Internet of Things," *Sensors (Basel, Switzerland)*, vol. 20, no. 9, pp. 1–23, 2020. doi:10.3390/s20092439.

[128] F. Alam, A. Almaghthawi, I. Katib, A. Albeshri, and R. Mehmood, "iResponse: An AI and IoT-Enabled Framework for Autonomous COVID-19 Pandemic Management," *Sustainability*, vol. 13, no. 7, p. 3797, 2021. doi:10.3390/su13073797.

[129] M. Dojat, F. Pachet, Z. Guessoum, D. Touchard, A. Harf, and L. Brochard, "Néoganesh: A working system for the automated control of assisted ventilation in icus," *Artif. Intell. Med*, vol. 11, no. 2, pp. 97–117,, 1997. doi:10.1016/s0933-3657(97)00025-0.

[130] P. Parthasarathy and S. Vivekanandan, "A typical iot architecture-based regular monitoring of arthritis disease using time wrapping algorithm," *Int. J. Comput. Appl*, vol. 42, no. 3, pp. 222–232,, 2020. doi:10.1080/1206212X.2018.1457471.

[131] M. Gaynor, "A user-centered, learning asthma smartphone application for patients and providers," *Learn. Heal. Syst*, no. December, pp. 1–12,, 2019. doi:10.1002/lrh2.10217.

[132] J. Car, W. Tan, Z. Huang, P. Sloot, and B. Franklin, "ehealth in the future of medications management: Personalisation, monitoring and adherence," *BMC Med*, vol. 15, no. 1, pp. 1–9,, 2017. doi:10.1186/s12916-017-0838-0.

[133] Y. Kim, "All-in-one, wireless, stretchable hybrid electronics for smart, connected, and ambulatory physiological monitoring," *Adv. Sci*, vol. 6, no. 17, 2019. doi:10.1002/advs.201900939.

[134] I. Bisio, C. Garibotto, F. Lavagetto, and A. Sciarrone, "Towards iot-based ehealth services: A smart prototype system for home rehabilitation," in *2019 IEEE Glob. Commun. Conf. GLOBECOM 2019 - Proc*, 2019. doi:10.1109/GLOBECOM38437.2019.9013194 pp. 1–6,.

[135] A. Nithya and A. Ranjani, "Monitoring health index of patient using iot in smart home environment," *Int. J. Adv. Sci. Technol*, vol. 28, no. 17, pp. 128–135,, 2019.

[136] W. Aman and F. Kausar, "Towards a gateway-based context-aware and self-adaptive security management model for iot-based ehealth systems," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 1, pp. 280–287,, 2019. doi:10.14569/IJACSA.2019.0100137.

[137] Q. Lohmeyer, A. Schneider, C. Jordi, J. Lange, and M. Meboldt, "Toward a new age of patient centricity? the application of eye-tracking to the development of connected self-injection systems," *Expert Opin. Drug Deliv*, vol. 16, no. 2, pp. 163–175,, 2019. doi:10.1080/17425247.2019.1563070.

[138] S. Moturi, S. Rao, and S. Vemuru, "Predictive analysis of imbalanced cardiovascular disease using smote," *Int. J. Adv. Sci. Technol*, vol. 29, no. 5, pp. 6301–6311,, 2020.

[139] G. Liu, C. Lustenberger, Y. Lo, W. Liu, Y. Sheu, and H. Wu, "Save muscle information–unfiltered eeg signal helps distinguish sleep stages," *Sensors (Switzerland)*, vol. 20, no. 7, pp. 1–12,, 2020. doi:10.1038/s41598-020-64083-4.

[140] E. Agliari, A. Barra, O. Barra, A. Fachechi, L. Vento, and L. Moretti, "Detecting cardiac pathologies via machine learning on heart-rate variability time series and related markers," *Sci. Rep*, vol. 10, no. 1, pp. 1–18,, 2020. doi:10.1038/s41598-020-64083-4.

[141] O. Debauche, S. Mahmoudi, P. Manneback, and A. Assila, "Fog iot for health: A new architecture for patients and elderly monitoring," *Procedia Comput. Sci*, vol. 160, pp. 289–297,, 2019. doi:10.1016/j.procs.2019.11.087.

[142] R. Harte, "Enhancing home health mobile phone app usability through general smartphone training: Usability and learnability case study," *J. Med. Internet Res*, vol. 20, no. 4, pp. 1–16,, 2018. doi:10.2196/humanfactors.7718.

[143] B. Bashar and M. Ismail, "Intelligent alarm system for hospitals using smartphone technology," *Telkomnika (Telecommunication Comput. Electron. Control*, vol. 18, no. 1, pp. 450–455,, 2020. doi:10.12928/TELKOMNIKA.V18I1.13262.

[144] M. Azad, J. Arshad, S. Mahmoud, K. Salah, and M. Imran, "A privacy-preserving framework for smart context-aware healthcare applications," *Trans. Emerg. Telecommun. Technol*, 2019. doi:10.1002/ett.3634.

[145] A. Ghosh, A. Raha, and A. Mukherjee, "Energy-efficient iot-health monitoring system using approximate computing," *Internet of Things*, vol. 9, pp. 100 166,, 2020. doi:10.1016/j.iot.2020.100166.

[146] M. Hossain, M. Rahman, and G. Muhammad, "Towards energy-aware cloud-oriented cyber-physical therapy system," *Futur. Gener. Comput. Syst*, 2020. doi:10.1016/j.future.2017.08.045.

[147] A. Martinez-Millana, C. Palao, C. Fernandez-Llatas, P. Carvalho, A. Bianchi, and V. Traver, "Integrated iot intelligent system for the automatic detection of cardiac variability," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf*, pp. 5798–5801,, 2018. doi:10.1109/EMBC.2018.8513638.

[148] A. Colomer, J. Igual, and V. Naranjo, "Detection of early signs of diabetic retinopathy based on textural and morphological information in fundus images," *Sensors (Switzerland)*, vol. 20, no. 4, 2020. doi:10.3390/s20041005.

[149] A. Follmann, M. Ohligs, N. Hochhausen, S. Beckers, R. Rossaint, and M. Czaplik, "Technical support by smart glasses during a mass casualty incident: A randomized controlled simulation trial on technically assisted triage and telemedical app use in disaster medicine," *J. Med. Internet Res*, vol. 21, no. 1, pp. 1–10,, 2019. doi:10.2196/11939.

[150] S. Karthikeyan, S. Srinivasan, J. Ali, and A. Veeraraghavan, "Smart summoning of ambulance during a vehicle accident," in *Proc. 2nd Int. Conf. Green Comput. Internet Things, ICGCIoT 2018*, 2018. doi:10.1109/ICGCIoT.2018.8752990 pp. 418–423,.

[151] Y. Luo, W. Li, and S. Qiu, "Anomaly detection based latency-aware energy consumption optimization for iot data-flow services," *Sensors (Switzerland)*, vol. 20, no. 1, 2020. doi:10.3390/s20010122.

[152] A. Gatouillat, B. Massot, Y. Badr, E. Sejdic, and C. Gehin, "Evaluation of a real-time low-power cardiorespiratory sensor for the iot," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf*, pp. 5382–5385,, 2018. doi:10.1109/EMBC.2018.8513550.

[153] C. Pagiatakis, D. Rivest-Hénault, D. Roy, F. Thibault, and D. Jiang, "Intelligent interaction interface for medical emergencies: Application to mobile hypoglycemia management," *Smart Health*, vol. 15, 2020. doi:10.1016/j.smhl.2019.100091.

[154] A. Mobashsher and A. Abbosh, "On-site rapid diagnosis of intracranial hematoma using portable multi-slice microwave imaging system," *Sci. Rep*, vol. 6, no. April, pp. 1–17,, 2016. doi:10.1038/srep37620.

[155] A. l'Aulnoit, "Development of a smart mobile data module for fetal monitoring in e-healthcare," *J. Med. Syst*, vol. 42, no. 5, pp. 1–7,, 2018. doi:10.1007/s10916-018-0938-1.

[156] L. Rachakonda, S. Mohanty, and E. Kougianos, "Ilog: An intelligent device for automatic food intake monitoring and stress detection in the iomt," *IEEE Trans. Consum. Electron*, vol. 66, no. 2, pp. 115–124,, 2020. doi:10.1109/TCE.2020.2976006.

[157] T. Fernández-Caramés, I. Froiz-Míguez, O. Blanco-Novoa, and P. Fraga-Lamas, "Enabling the internet of mobile crowdsourcing health things: A mobile fog computing, blockchain and iot based continuous glucose monitoring system for diabetes mellitus research and care," *Sensors*, vol. 19, no. 15, pp. 1–24,, 2019. doi:10.3390/s19153319.

[158] M. Al-khafajiy, H. Kolivand, T. Baker, D. Tully, and A. Waraich, "Smart hospital emergency system," *Multimed. Tools Appl*, vol. 78, no. 14, pp. 20 087–20 111,, 2019. doi:10.1007/s11042-019-7274-4.

[159] R. D. A. Lauraitis, R. M. unas and T. Krilavi ˇ, "A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment," *Sensors*, 2020.

[160] K. Cho, "Detecting patient deterioration using artificial intelligence in a rapid response system," *Crit. Care Med*, 2020. doi:10.1097/CCM.0000000000004236.

[161] F. Trenta, S. Conoci, F. Rundo, and S. Battiato, "Advanced motion-tracking system with multi-layers deep learning framework for innovative car-driver drowsiness monitoring," in *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, 2019. doi:10.1109/FG.2019.8756566.

[162] W. Noel, R. Bosc, S. Jabbour, E. Kechichian, B. Hersant, and J.-P. Meningaud, "Smartphone-based patient education in plastic surgery," *Ann. Plast. Surg*, vol. 79, no. 6, pp. 529–531,, 2017. doi:10.1097/SAP.0000000000001241.

[163] G. K and K. K, "Application of IoT in Predictive Health Analysis – A Review of Literature," *International Journal of Management, Technology, and Social Sciences (IJMTS)*, vol. 5, p. 185–214, 2020. doi:10.5281/zenodo.3821147.

[164] C. Schuh, J. Bruin, and W. Seeling, "Clinical decision support systems at the vienna general hospital using arden syntax: Design, implementation, and integration," *Artif. Intell. Med*, 2018. doi:10.1016/j.artmed.2015.11.002.

[165] R. Miotto, M. Danieletto, J. R. Scelza, B. A. Kidd, and J. T. Dudley, "Reflecting health: smart mirrors for personalized medicine," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–7, 2018. doi:10.1038/s41746-018-0068-7.

[166] M. Abdel-Basset, G. Manogaran, A. Gamal, and V. Chang, "A novel intelligent medical decision support model based on soft computing and iot," *IEEE Internet Things J*, vol. 7, no. 5, pp. 4160–4170,, 2020. doi:10.1109/JIOT.2019.2931647.

[167] M. Fernandes, S. Vieira, F. Leite, C. Palos, S. Finkelstein, and J. Sousa, "Clinical decision support systems for triage in the emergency department using intelligent systems: a review," *Artificial Intelligence in Medicine*, 2020. doi:10.1016/j.artmed.2019.101762.

[168] M. Aldeer, M. Alaziz, J. Ortiz, R. Howard, and R. Martin, "A sensing-based framework for medication compliance monitoring," in *DFHS 2019 - Proc. 1st ACM Work. Device-Free Hum. Sens*, 2019. doi:10.1145/3360773.3360886 p. 52–56.

[169] V. Adjiski, Z. Despodov, D. Mirakovski, and D. Serafimovski, "System architecture to bring smart personal protective equipment wearables and sensors to transform safety at work in the underground mining industry," *MGPB*, vol. 34, no. 1, pp. 37–44,, 2019. doi:10.17794/rgn.2019.1.4.

[170] I. Kagan, M. Hellerman-Itzhaki, I. Neuman, Y. Glass, and P. Singer, "Reflux events detected by multichannel bioimpedance smart feeding tube during high flow nasal cannula oxygen therapy and enteral feeding: First case report," *J. Crit. Care*, vol. 60, pp. 226–229,, 2020. doi:10.1016/j.jcrc.2020.08.005.

[171] R. Miotto, M. Danieletto, J. R. Scelza, B. A. Kidd, and J. T. Dudley, "Reflecting health: smart mirrors for personalized medicine," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–7, 2018. doi:10.1038/s41746-018-0068-7.

[172] S. Pai, S. Hui, R. Isserlin, M. A. Shah, H. Kaka, and G. D. Bader, "netDx: interpretable patient classification using integrated patient similarity networks," *Molecular Systems Biology*, vol. 15, no. 3, 2019. doi:10.15252/msb.20188497.

[173] S. Boriah, V. Chandola, and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," *Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics 130*, vol. 1, pp. 243–254, 2008. doi:10.1137/1.9781611972788.22.

[174] M. Verleysen, D. François, G. Simon, and V. Wertz, "On the effects of dimensionality on data analysis with neural networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2687, pp. 105–112, 2003. doi:10.1007/3-540-44869-1_14.

[175] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in *Database Theory — ICDT 2001*, J. den Bussche and V. Vianu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. doi:10.1007/3-540-44503-X_27. ISBN 978-3-540-44503-6 pp. 420–434.

[176] J. Han, M. Kamber, and J. Pei, "Getting to Know Your Data," *Data Mining*, pp. 39–82, 2012. doi:10.1016/B978-0-12-381479-1.00002-2.

[177] D. B. Bisandu, R. Prasad, . Musa, M. Liman, D. Bala Bisandu, and M. M. Liman, "Data clustering using efficient similarity measures," *Journal of Statistics and Management Systems*, vol. 22, no. 5, pp. 901–922, 2019. doi:10.1080/09720510.2019.1565443.

[178] M. Panahiazar, V. Taslimitehrani, N. L. Pereira, M.D, and J. Pathak, "Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics," *Studies in Health Technology and Informatics*, vol. 210, p. 369–373, 2015. doi:10.3233/978-1-61499-512-8-369.

[179] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang, "Measuring patient similarities via a deep architecture with medical concept embedding," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 749–758, 2017. doi:10.1109/ICDM.2016.0086.

[180] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to Diagnose with LSTM Recurrent Neural Networks," pp. 1–18, 2015, [Accessed: 2023-09-12]. [Online]. Available: http://arxiv.org/abs/1511.03677.

[181] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, A. Zhang, and J. Gao, "Personalized disease prediction using a CNN-based similarity learning method," *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, vol. 2017-Janua, pp. 811–816, 2017. doi:10.1109/BIBM.2017.8217759.

[182] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE Transactions on Nanobioscience*, vol. 17, no. 3, pp. 219–227, 2018. doi:10.1109/TNB.2018.2837622.

[183] C. Combi, M. Gozzi, J. M. Juarez, R. Marin, and B. Oliboni, "Querying clinical workflows by temporal similarity," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, R. Bellazzi, A. Abu-Hanna, and J. Hunter, Eds., vol. 4594 LNAI. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. doi:10.1007/978-3-540-73599-1_63. ISBN 3540735984. ISSN 16113349 pp. 469–478.

[184] A. Gottlieb, G. Y. Stein, E. Ruppin, R. B. Altman, and R. Sharan, "A method for inferring medical diagnoses from patient similarities," *BMC Medicine*, vol. 11, no. 1, 2013. doi:10.1186/1741-7015-11-194.

[185] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PLoS ONE*, vol. 10, no. 5, pp. 1–13, 2015. doi:10.1371/journal.pone.0127428.

[186] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Scientific Reports*, vol. 6, no. April, pp. 1–10, 2016. doi:10.1038/srep26094.

[187] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014. doi:10.1038/nmeth.2810.

[188] K. Ng, J. Sun, J. Hu, and F. Wang, "Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity." *AMIA Joint Summits on Translational Science proceedings*, vol. 2015, pp. 132–6, Mar 2015.

[189] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: A patient-centered framework," *Journal of General Internal Medicine*, vol. 28, no. SUPPL.3, pp. 660–665, 2013. doi:10.1007/s11606-013-2455-8.

[190] I. Song and N. V. Marsh, "Anonymous indexing of health conditions for a similarity measure," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 737–744, 2012. doi:10.1109/TITB.2012.2194717.

[191] L. Chan, T. Chan, L. Cheng, and W. Mak, "Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy," pp. 467–470, Dec 2010. doi:10.1109/BIBMW.2010.5703846.

[192] D. Girardi, S. Wartner, G. Halmerbauer, M. Ehrenmüller, H. Kosorus, and S. Dreiseitl, "Using concept hierarchies to improve calculation of patient similarity," *Journal of Biomedical Informatics*, vol. 63, pp. 66–73, 2016. doi:10.1016/j.jbi.2016.07.021.

[193] D. Heckerman, "Probabilistic similarity networks," *Networks*, vol. 20, no. 5, pp. 607–636, 1990. doi:10.1002/net.3230200508.

[194] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Update on the Pathfinder Project," *Annual Symposium on Computer Application in Medical Care*, no. November 1989, pp. 203–207, 1989, [Accessed: 2023-01-3]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245693/pdf/procascamc00017-0215.pdf.

[195] Y. Wang, Y. Tian, L. L. Tian, Y. M. Qian, and J. S. Li, "An Electronic Medical Record System with Treatment Recommendations Based on Patient Similarity," *Journal of Medical Systems*, vol. 39, no. 5, 2015. doi:10.1007/s10916-015-0237-z.

[196] F. S. Roque, P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Søeby, S. Bredkjær, A. Juul, T. Werge, L. J. Jensen, and S. Brunak, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS Computational Biology*, vol. 7, no. 8, 2011. doi:10.1371/journal.pcbi.1002141.

[197] K. Lage, E. O. Karlberg, Z. M. Størling, P. Ólason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007. doi:10.1038/nbt1295.

[198] N. D. Seligson, J. L. Warner, W. S. Dalton, D. Martin, R. S. Miller, D. Patt, K. L. Kehl, M. B. Palchuk, G. Alterovitz, L. K. Wiley, M. Huang, F. Shen, Y. Wang, K. A. Nguyen, A. F. Wong, F. Meric-Bernstam, E. V. Bernstam, and J. L. Chen, "Recommendations for patient similarity classes: results of the AMIA 2019 workshop on defining patient similarity," *Journal of the American Medical Informatics Association*, pp. 1–5, 2020. doi:10.1093/jamia/ocaa159.

[199] A. Tashkandi, I. Wiese, and L. Wiese, "Efficient In-Database Patient Similarity Analysis for Personalized Medical Decision Support Systems," *Big Data Research*, vol. 13, pp. 52–64, 2018. doi:10.1016/j.bdr.2018.05.001.

[200] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining Drug and Gene Similarity Measures for Drug-Target Elucidation," *Journal of Computational Biology*, vol. 18, no. 2, pp. 133–145, 2011. doi:10.1089/cmb.2010.0213.

[201] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson, "Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies," *American Journal of Human Genetics*, vol. 85, no. 4, pp. 457–464, 2009. doi:10.1016/j.ajhg.2009.09.003.

[202] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, and X. Jiang, "Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis," *JMIR Medical Informatics*, vol. 6, no. 2, 4 2018. doi:10.2196/medinform.7744.

[203] S. Koks, R. W. Williams, J. Quinn, F. Farzaneh, N. Conran, S. J. Tsai, G. Awandare, and S. R. Goodman, "Highlight article: COVID-19: Time for precision epidemiology," *Experimental Biology and Medicine*, vol. 245, no. 8, pp. 677–679, 2020. doi:10.1177/1535370220919349.

[204] P. Hartono, "Similarity maps and pairwise predictions for transmission dynamics of COVID-19 with neural networks," *Informatics in Medicine Unlocked*, vol. 20, p. 100386, 2020. doi:10.1016/j.imu.2020.100386.

[205] K. Gao, D. D. Nguyen, R. Wang, and G.-W. Wei, "Machine intelligence design of 2019-nCoV drugs." *bioRxiv : the preprint server for biology*, p. 2020.01.30.927889, 2020. doi:10.1101/2020.01.30.927889.

[206] M. P. Shahri, K. Lyon, J. Schearer, and I. Kahanda, "DeepPPPred: An Ensemble of BERT, CNN, and RNN for Classifying Co-mentions of Proteins and Phenotypes," *bioRxiv*, p. 2020.09.18.304329, 2020. doi:10.1101/2020.09.18.304329.

[207] Y. Xiong, S. Chen, H. Qin, H. Cao, Y. Shen, X. Wang, Q. Chen, J. Yan, and B. Tang, "Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity," *BMC Medical Informatics and Decision Making*, vol. 20, no. Suppl 1, pp. 1–7, 2020. doi:10.1186/s12911-020-1045-z.

[208] M. Žitnik and B. Zupan, "Data fusion by matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 41–53, 2015. doi:10.1109/TPAMI.2014.2343973.

[209] J. Ni, J. Liu, C. Zhang, D. Ye, and Z. Ma, “Fine-grained Patient Similarity Measuring using Deep Metric Learning,” no. November 2017, pp. 1189–1198, 2017. doi:10.1145/3132847.3133022.

[210] R. Y. WANG and D. M. STRONG, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996. doi:10.1080/07421222.1996.11518099.

[211] C. L. Pritchard, “Data Quality Assessment,” *Risk Management*, pp. 285–292, 2020. doi:10.1201/9781439842201-34.

[212] L. Sebastian-Coleman, “Section 3. Data Assessment Scenarios,” *Measuring Data Quality for Ongoing Improvement*, pp. 93–96, 2013. doi:10.1016/b978-0-12-397033-6.00037-7.

[213] J. Patterson and A. Gibson, *Deep learning: A practitioner’s approach*. " O’Reilly Media, Inc.", 2017. ISBN 1491914238.

[214] D. Loshin, “Dimensions of Data Quality,” *The Practitioner’s Guide to Data Quality Improvement*, pp. 129–146, 2011. doi:10.1016/b978-0-12-373717-5.00008-7.

[215] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Computing Surveys*, vol. 41, no. 3, 2009. doi:10.1145/1541880.1541883.

[216] P. Woodall, M. Oberhofer, and A. Borek, “A classification of data quality assessment and improvement methods,” *International Journal of Information Quality*, vol. 3, no. 4, pp. 298–321, 2014. doi:10.1504/IJIQ.2014.068656.

[217] J. Debattista, S. Auer, and C. Lange, “Luzzu - A methodology and framework for linked data quality assessment,” *Journal of Data and Information Quality*, vol. 8, no. 1, pp. 1–32, 2016. doi:10.1145/2992786.

[218] L. Ehrlinger, V. Haunschmid, D. Palazzini, and C. Lettner, “A DaQL to Monitor Data Quality in Machine Learning Applications,” in *Lecture Notes in Computer Science*, 2019. doi:10.1007/978-3-030-27615-7_17.

[219] C. Lettner, R. Stumptner, W. Fragner, F. Rauchenzauner, and L. Ehrlinger, "DaQL 2.0: Measure Data Quality based on Entity Models," *Procedia Computer Science*, vol. 180, no. 2019, pp. 772–777, 2021. doi:10.1016/j.procs.2021.01.327.

[220] W. Dai, K. Yoshigoe, and W. Parsley, "Improving Data Quality Through Deep Learning and Statistical Models," in *Information Technology - New Generations*, S. Latifi, Ed. Cham: Springer International Publishing, 2018. doi:10.1007/978-3-319-54978-1_66. ISBN 978-3-319-54978-1 pp. 515–522.

[221] D. Loshin, "Data Quality and MDM," *Master Data Management*, pp. 87–103, 2009. doi:10.1016/b978-0-12-374225-4.00005-9.

[222] S. Shekhar and A. Gokhale, "Dynamic resource management across cloud-edge resources for performance-sensitive applications," *Proceedings - 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017*, pp. 707–710, 2017. doi:10.1109/CCGRID.2017.120.

[223] "Covid-19: China's digital health strategies against the global pandemic," *ITU News*, 2020-04, [Accessed: 2023-05-29]. [Online]. Available: https://www.itu.int/hub/2020/05/covid-19-chinas-digital-health-strategies-against-the-global-pandemic/.

[224] Y. Zhan, P. Li, and S. Guo, "Experience-Driven Computational Resource Allocation of Federated Learning by Deep Reinforcement Learning," *Proceedings - 2020 IEEE 34th International Parallel and Distributed Processing Symposium, IPDPS 2020*, pp. 234–243, 2020. doi:10.1109/IPDPS47924.2020.00033.

[225] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing Federated Learning on Non-IID Data with Reinforcement Learning," *Proceedings - IEEE INFOCOM*, vol. 2020-July, pp. 1698–1707, 2020. doi:10.1109/INFOCOM41043.2020.9155494.

[226] W. Y. B. Lim, J. S. Ng, Z. Xiong, D. Niyato, C. Miao, and D. I. Kim, "Dynamic Edge Association and Resource Allocation in Self-Organizing Hierarchical Federated Learning Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3640–3653, 2021. doi:10.1109/JSAC.2021.3118401.

[227] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair Resource Allocation in Federated Learning," pp. 1–27, 2019. doi:10.48550/arXiv.1905.10497.

[228] W. Xia, T. Q. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit-based client scheduling for federated learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7108–7123, 2020. doi:10.1109/TWC.2020.3008091.

[229] V. W. Anelli, Y. Deldjoo, T. Di Noia, and A. Ferrara, "Towards Effective Device-Aware Federated Learning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11946 LNAI, pp. 477–491, 2019. doi:10.1007/978-3-030-35166-3_34.

[230] Z. Chai, H. Fayyaz, Z. Fayyaz, A. Anwar, Y. Zhou, N. Baracaldo, H. Ludwig, and Y. Cheng, "Towards taming the resource and data heterogeneity in federated learning," in *2019 USENIX Conference on Operational Machine Learning (OpML 19)*. Santa Clara, CA: USENIX Association, May 2019. ISBN 978-1-939133-00-7 pp. 19–21. [Online]. Available: https://www.usenix.org/conference/opml19/presentation/chai.

[231] Z. Chai, A. Ali, S. Zawad, S. Truex, A. Anwar, N. Baracaldo, Y. Zhou, H. Ludwig, F. Yan, and Y. Cheng, "TiFL: A Tier-based Federated Learning System," *HPDC 2020 - Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, no. 1, pp. 125–136, 2020. doi:10.1145/3369583.3392686.

[232] L. U. Khan, W. Saad, Z. Han, and C. S. Hong, "Dispersed Federated Learning: Vision, Taxonomy, and Future Directions," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 192–198, 2021. doi:10.1109/MWC.011.2100003.

[233] L. L. Pilla, "Optimal task assignment for heterogeneous federated learning devices," *Proceedings - 2021 IEEE 35th International Parallel and Distributed Processing Symposium, IPDPS 2021*, pp. 661–670, 2021. doi:10.1109/IPDPS49936.2021.00074.

[234] T. Huang, W. Lin, W. Wu, L. He, K. Li, and A. Y. Zomaya, "An Efficiency-Boosting Client Selection Scheme for Federated Learning with Fairness Guarantee," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1552–1564, 2021. doi:10.1109/TPDS.2020.3040887.

[235] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," pp. 1–8, 2021. doi:10.1073/pnas.2024789118.

[236] J. Jeon, S. Park, M. Choi, J. Kim, Y. B. Kwon, and S. Cho, "Optimal user selection for high-performance and stabilized energy-efficient federated learning platforms," *Electronics (Switzerland)*, vol. 9, no. 9, pp. 1–17, 2020. doi:10.3390/electronics9091359.

[237] A. Albaseer, M. Abdallah, A. Al-Fuqaha, and A. Erbad, "Client Selection Approach in Support of Clustered Federated Learning over Wireless Edge Networks," *2021 IEEE Global Communications Conference, GLOBECOM 2021 - Proceedings*, 2021. doi:10.1109/GLOBECOM46510.2021.9685938.

[238] J. Xu and H. Wang, "Client Selection and Bandwidth Allocation in Wireless Federated Learning Networks: A Long-Term Perspective," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1188–1200, 2021. doi:10.1109/TWC.2020.3031503.

[239] P. Nair, P. E. Antoniou, E. J. Pino, and G. Fico, "Editorial: Highlights in connected health 2021/22," *Frontiers in Digital Health*, vol. 4, no. November, pp. 1–3, 2022. doi:10.3389/fdgth.2022.1066860.

[240] M. A. Serhani, A. N. Navaz, H. Al Ashwal, and N. Al Qirim, "Ecg-based arrhythmia classification & clinical suggestions: An incremental approach of hyperparameter tuning," *ACM International Conference Proceeding Series*, pp. 13–19, 2020. doi:10.1145/3419604.3419787.

[241] H. Wang, M. Daneshmand, and H. Fang, "Artificial Intelligence (AI) driven wireless body area networks: Challenges and directions," *Proceedings - IEEE International Conference on Industrial Internet Cloud, ICII 2019*, no. Icii, pp. 428–429, 2019. doi:10.1109/ICII.2019.00079.

[242] H. Cheng, P.-N. Tan, J. Gao, and J. Scripps, "Multistep-ahead time series prediction," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2006. doi:10.1007/11731139_89 pp. 765–774.

[243] R. Isele and C. Bizer, "Learning linkage rules using genetic programming," in *Proceedings of the 6th International Conference on Ontology Matching-Volume 814*. CEUR-WS. org, 2011. doi:10.48550/arXiv.1208.0291 pp. 13–24.

[244] H. Smith and D. P. Marketing, "Data Profiling : The First Step to Big Data Quality," Tech. Rep.

[245] V. Bolón-Canedo, K. Sechidis, N. Sánchez-Maroño, A. Alonso-Betanzos, and G. Brown, "Insights into distributed feature ranking," *Information Sciences*, vol. 496, pp. 378–398, 2019. doi:10.1016/j.ins.2018.09.045.

[246] Y. Xu, L. Ma, Y. Fan, Y. Y. Chen, K. Ma, J. Yang, X. Yang, Y. Y. Chen, C. Shu, Z. Fan, J. Gan, X. Zou, R. Huang, C. Zhang, X. Liu, D. Tu, C. Xu, W. Zhang, D. Yang, M.-W. Wang, X. Wang, X. Xie, H. Leng, N. Holalkere, N. J. Halin, I. R. Kamel, J. Wu, X. Peng, X. Wang, J. Shao, P. Mongkolwat, J. Zhang, D. L. Rubin, G. Wang, C. Zheng, Z. Li, X. Bai, T. Xia, F. Yang, Y. Y. Chen, K. Ma, J. Yang, X. Yang, Y. Y. Chen, C. Shu, Z. Fan, J. Gan, X. Zou, R. Huang, C. Zhang, X. Liu, D. Tu, C. Xu, W. Zhang, D. Yang, M.-W. Wang, X. Wang, X. Xie, H. Leng, N. Holalkere, N. J. Halin, I. R. Kamel, J. Wu, X. Peng, X. Wang, J. Shao, P. Mongkolwat, J. Zhang, D. L. Rubin, G. Wang, C. Zheng, Z. Li, X. Bai, and T. Xia, "A collaborative online AI engine for CT-based COVID-19 diagnosis." *medRxiv : the preprint server for health sciences*, p. 2020.05.10.20096073, 2020. doi:10.1101/2020.05.10.20096073.

[247] "Framingham Heart Study," [Accessed: 2023-06-9]. [Online]. Available: https://www.framinghamheartstudy.org/participants/participant-cohorts/.

[248] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[249] "Weighted Scoring | Definition and Overview," [Accessed: 2023-04-26]. [Online]. Available: https://www.productplan.com/glossary/weighted-scoring/.

[250] A. Castiglioni, "Dimensionality reduction with autoencoders versus pca | towards data science," [Accessed: 2023-06-24]. [Online]. Available: https://towardsdatascience.com/dimensionality-reduction-with-autoencoders-versus-pca-f47666f80743.

[251] Z. Song, "Performance of Autoencoder with Bi-Directional Long-Short Term Memory Network in Gestures Unit Segmentation," vol. 1, no. 1989, pp. 1–6, 2018.

[252] J. Chen, "The effect of an auto-encoder on the accuracy of a convolutional neural network classification task," *Res. Sch. Comput. Sci, Aust. Nat. Univ.*, p. 1–8, 2018, [Accessed: 2022-04-2]. [Online]. Available: https://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_166.pdf.

[253] D. Ayres-de campos, J. Bernardes, A. Garrido, J. Marques-de sá, and L. Pereira-leite, "SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms," *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 9, no. 5, pp. 311–318, 2000. doi:10.3109/14767050009053454.

List of Other Publications

Journal Publications

I. A. N. Navaz, H. T. El Kassabi, M. A. Serhani, and E. S. Barka, "Resource-aware Federated Hybrid Profiling for Edge Node Selection in Federated Patient Similarity Network," *MDPI Applied Sciences*, vol. 13, no. 23, p. 13114, 2023, DOI: 10.3390/app132413114.

II. A. N. Navaz, M. A. Serhani, H. T. El Kassabi, and I. Taleb, "Empowering Patient Similarity Networks through Innovative Data- Quality-Aware Federated Profiling," *Sensors (Basel, Switzerland)*, vol. 23, no. 14, pp. 1–32, 2023, ISSN: 14248220. DOI: 10.3390/s23146443.

III. A. N. Navaz, H. T. El-kassabi, M. A. Serhani, A. Oulhaj, and K. Khalil, "A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine," *Journal of Personalized Medicine*, 2022. 10.3390/jpm12050768.

IV. H. Ismail, M. A. Serhani, N. Hussien, R. Elabyad, and A. Navaz, "Public wellbeing analytics framework using social media chatter data," *Social Network Analysis and Mining*, 2022. DOI: 10.1007/s13278-022-00987-5.

V. A. N. Navaz, M. A. Serhani, H. T. El Kassabi, N. Al-Qirim, and H. Ismail, "Trends, Technologies, and Key Challenges in Smart and Connected Healthcare," *IEEE Access*, 2021, ISSN: 21693536. DOI: 10.1109/ACCESS.2021.3079217.

VI. T. Habuza, A. N. Navaz, F. Hashim, et al., "AI applications in robotics, diagnostic image analysis and precision medicine: Current limitations, future trends, guidelines on CAD systems for medicine," *Informatics in Medicine Unlocked (IMU)*, vol. 24, 2021. DOI: 10.1016/j.imu.2021.100596.

VII. M. A. Serhani, H. T. El Kassabi, H. Ismail, and A. N. Navaz, "ECG monitoring systems: Review, architecture, processes, and key challenges," *Sensors (Switzerland)*, 2020. DOI:10.3390/s20061796.

VIII. H. T. El-Kassabi, M. Adel Serhani, R. Dssouli, and A. N. Navaz, "Trust enforcement through self-adapting cloud workflow orchestration," *Future Generation Computer Systems*, 2019. DOI: 10.1016/j.future.2019.03.004.

IX. A. N. Navaz, M. A. Serhani, N. Al-Qirim, and M. Gergely, "Towards an efficient and Energy-Aware mobile big health data architecture," *Computer Methods and Programs in Biomedicine*, 2018. DOI: 10.1016/j.cmpb.2018.10.008.

X. M. Masud, M. Adel Serhani, and A. Navaz, "Resource-Aware Mobile-Based Health Monitoring," *IEEE Journal of Biomedical and Health Informatics*, 2017. DOI:10.1109/JBHI.2016.2525006.

XI. M. Serhani, M. Menshawy, A. Benharref, S. Harous, and A. Navaz, "New algorithms for processing time-series big EEG data within mobile health monitoring systems," *Computer Methods and Programs in Biomedicine*, 2017. DOI: 10.1016/j.cmpb.2017.07.007.

XII. M. A. Serhani, A. Benharref, and A. R. Nujum, "An adaptive expert system for automated advices generation-based semicontinuous M-health monitoring," *Brain Informatics and Health*, BIH 2014,DOI: 10.1007/978-3-319-09891-$3_3$6.

Conference Proceedings

XIII. A. N. Navaz, M. A. Serhani, and H. El Kassabi, "Federated Quality Profiling : A quality evaluation of patient monitoring at the Edge," *IEEE*, 2022. DOI: 10.1109/IWCMC55113.2022.9825083.

XIV. H. T. El Kassabi, M. Adel Serhani, A. N. Navaz, and S. Ouhbi, "Federated Patient Similarity Network for Data-Driven Diagnosis of COVID-19 Patients," in *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, 2021. DOI: 10.1109/AICCSA53542.2021.9686875.

XV. M. A. Serhani, A. N. Navaz, H. Al Ashwal, and N. Al Qirim, "ECG based arrhythmia classification clinical suggestions: An incremental approach of hyperparameter tuning," 2020.

XVI. A. N. Navaz, S. M. Adel, and S. S. Mathew, “Facial image preprocessing and emotion classification: A deep learning approach,” 2019. DOI: 10.1109/AICCSA47632.2019.9035268.

XVII. A. N. Navaz, S. Harous, M. A. Serhani, and I. Taleb, “Real-Time Data Streaming Algorithms and Processing Technologies: A Survey,” *IEEE*, 2019. DOI: 10.1109/ICCIKE47802.2019.9004318.

XVIII. M. Serhani, H. El Kassabi, N. Al Qirim, and A. Navaz, “Towards a Multi-model CloudWorkflow Resource Monitoring, Adaptation, and Prediction,” in *17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering*, 2018. DOI: 10.1109/TrustCom/BigDataSE.2018.00265.

XIX. A. Navaz, E. Mohammed, M. Serhani, and N. Zaki, in *Proceedings of the 2016 12th International Conference on Innovations in Information Technology*, IIT 2016. DOI:10.1109/INNOVATIONS.2016.7880045.

XX. M. Serhani, M. El Menshawy, A. Benharref, and A. Navaz, “Real time EEG compression for energy-aware continous mobile monitoring,” in *Proceedings of the International Conference on Microelectronics*, ICM, 2016. DOI: 10.1109/ICM.2015.7438046.

XXI. M. A. Serhani, H. T. El Kassabi, I. Taleb, and A. Nujum, “An hybrid approach to quality evaluation across big data value chain,” 2016. DOI: 10.1109/BigDataCongress.2016.65.

XXII. A. Benharref, M. A. Serhani, and A. R. Nujum, “Closing the loop from continuous M-health monitoring to fuzzy logic-based optimized recommendations,” IEEE, 2014. DOI: 10.1109/EMBC.2014.6944179.

XXIII. M. A. Serhani, A. Benharref, and A. R. Nujum, “Intelligent remote health monitoring using evident-based DSS for automated assistance,” IEEE, 2014. DOI: 10.1109/EMBC.2014.6944173.

UAEU جامعة الإمارات العربية المتحدة United Arab Emirates University

UAE UNIVERSITY DOCTORATE DISSERTATION NO. 2023: 66

This research presents a multi-model approach that includes a multi-dimensional Patient Similarity Network (PSN) Fusion model, a Data Quality Management model inspired by Federated Learning, and a PSN Resource Optimization Model. These innovations collectively address data heterogeneity, enhance data quality, and optimize resource utilization to improve healthcare outcomes in Smart and Connected Health (SCH).

www.uaeu.ac.ae

Alramzana Nujum Navaz received her PhD from the Department of Computer Science and Software Engineering, College of Information Technology at UAE University, UAE.

Online publication of dissertation: https://scholarworks.uaeu.ac.ae/etds/